Volume II

Appendices to the Strategic Evolution of Earth Science Enterprise (ESE) Data Systems (SEEDS) Formulation Team Final Recommendations Report

June 2003

Contents

About This Document

The material in this volume is meant to complement the contents of Volume I – Strategic Evolution of Earth Science Enterprise (ESE) Data Systems (SEEDS) Draft Recommendations, December 2002. The documents herein are detailed results of the work done by some of the seven study teams of the SEEDS Formulation Team, including methodology, data, statistical analyses, and other important background information. If the reader wishes to see only the draft recommendations, s/he is encouraged to read Volume I first. As these are final reports of studies, any comments should be directed to that team's study lead.

Appendix A – Levels of Service and Cost Estimation

Contents of Appendix A

SEEDS

Working Paper One:

Study Overview and Technical Approach

April 24, 2002

G. Hunolt, SGT, Inc.

Outline

Introduction

This working paper is the first of a set of papers that describes the SEEDS (Strategic Evolution of Earth Science Enterprise Data Systems) Levels of Service (LOS) / Cost Estimation (LOS/CE) study. The study goal is to develop a cost estimation model and coupled requirements and levels of services to support the SEEDS Formulation Team in estimating the life cycle costs of future ESE data service providers and supporting systems, where 'data service provider' is used as a generic term for any data/information related activity. The set of working papers is intended to serve as a vehicle for coordinating work on the project, obtaining feedback and guidance from ESDIS SOO and the user community, and as embryos of reports that will be produced as the task proceeds.

As working papers, each version of each paper that appears represents a snapshot in time, with the work in various stages of completion. As work progresses the content (and sometimes the organization) of the working papers will change reflecting progress made, responses to feedback and guidance received, etc.

This first working paper of the set provides an overview of the LOS/CE study, a roadmap to the full set of working papers, a statement of the objectives of the study and an outline of the technical approach being taken to meet the objectives of the levels of service and cost estimation phases of the study. It constitutes a high level plan for the study. The initial version of this paper is focused on the work of the study through June, 2002. A major update to this paper will be provided by June 30, 2002, which will address progress and plans for the next year.

This paper introduces the set of working papers in Section 2. Section 3 states the overall objectives of the LOS/CE Study. Section 4 presents an overview and high level schedule for the study. Section 5 presents the technical approach for the SGT portion of the effort, including some related notes and assumptions, and the approaches to be taken to the requirements analysis and cost model development, consistent with the SGT task plan submitted to ESDIS SOO in October, 2001.

Note: As of late November, 2001, SEEDS replaces the term 'NewDISS' under which the Formulation Team had begun work. The new term is intended to emphasize the Earth Science Enterprise's (ESE's) evolutionary approach. The term 'NewDISS' will be retained when it refers to NewDISS documents that predate the change in terminology. Similarly, the term 'data service provider' was adopted as the generic name for an ESE activity that provides any form of data and/or information management and user services, replacing the term 'data center', which will be used only in the more conventional sense as a type of data service provider.

Roadmap to the Set of Working Papers

This section describes the working papers that together describe the LOS/CE Study.

The initial version of the working papers is a decomposition of the previous overall working paper "SEEDS Requirements / LOS & Cost Model Working Paper - New Year's Draft", January 16, 2002, with updates made per the results of the February 5-7, 2002, SEEDS Community Workshop. As an aid to readers of the original document, the roadmap in the initial edition of this paper will contain references to the section numbers in the January 16 working paper from which material was taken.

Although the papers are numbered, except for Working Paper 1 the papers are not intended to be regarded as a sequence, but rather as parallel, and they will refer to each other freely. Their development and updating will be asynchronous, reflecting progress on the study as it occurs.

Working Paper 1 - Project Overview and Technical Approach

The first working paper of the set provides an overview of the SEEDS Levels of Service / Cost Estimation Study, a roadmap to the full set of working papers, and a discussion of the technical approach to the requirements analysis and cost estimation phases of the study, constituting a high level plan for the work to be done. A major update to this working paper will be provided by June 30, 2002. (Contains material from sections 1.1, 3.1, 3.2, 3.3 of January 16, 2002 working paper).

Working Paper 2 - Cost Estimation by Analogy Model

This working paper describes the cost estimation by analogy model that is being developed for this study. This paper will evolve extensively as the work progresses. Its initial focus is on a conceptual description of the model and how it and the cost estimating relationships it uses are expected to develop, scenarios showing how the model will be used, goals and plans for the model prototype planned for June 2002, and the plan for progressively more detailed documentation of computational processes used by the model as it develops. (Contains material from sections 2.1, 2.2, 2.3, 2.4, and 2.5 of the January 16, 2002 working paper.)

Working Paper 3 - Data Service Provider Reference Model - Functional Areas

This working paper describes the concepts involved in the Data Service Provider Reference Model, and describes the functional areas / areas of cost comprising the model. The paper reflects the results of the February, 2002, SEEDS Community Workshop, including drawing on material from white papers submitted by workshop attendees. (Contains material from sections 3.1 and 4.1 of the January 16, 2002 working paper.)

Working Paper 4 - Data Service Provider Reference Model - Model Parameters

This working paper contains definitions of the parameters that are inputs, outputs, and intermediate parameters used by the cost estimation by analogy model, including those that are elements of the comparables database. It constitutes a data dictionary for the model and database. (Contains material from sections 4.2 through 4.5 of the January 16, 2002, working paper.)

Working Paper 5 - Data Service Provider Reference Model - Requirements / Levels of Service

This working paper describes a general set of requirements and levels of service mapped to the functional areas of the Data Service Provider Reference Model. This paper will be maintained and updated as needed through the life of the project. This paper reflects the results of the February, 2002, Community Workshop, draws on white papers submitted by workshop attendees, and includes a new user-oriented view of levels of service. (Contains material from sections 5.1 through 5.11 of the January 16, 2002, working paper.)

Working Paper 6 - ESE Logical Data Service Provider Types

This working paper describes an open set of logical ESE data service provider types, each essentially a group of functions clustered around a different ESE role or mission as an organizing principle. The paper describes how these logical or conceptual provider types relate to physical entities, e.g. real-world data centers that, given their responsibilities within the ESE program, might embody the functionality of several different provider types. The paper describes how the provider types would be used in ESE architecture studies. The paper reflects the results of the February, 2002, Community Workshop, and draws on white papers submitted by workshop attendees. (Contains material from sections 6.1 - 6.8 of the January 16 working paper.)

Working Paper 7 - Comparables Database

This working paper provides an overview of the Comparables Database, comprising information obtained from existing ESE data activities and other data centers. It includes the database schema or template. It reports on which data centers have provided information to be added to the database, allowing a reader to track the development of the database as the information collection effort proceeds and the paper is updated. The paper does not contain the actual information provided by the sites.

LOS and Cost Estimation Study Objectives

Key facets of the SEEDS Formulation study will be to establish the minimum levels of service that ESE data service providers will be required to provide for the user community and to provide the capability to estimate costs for ESE data service providers to provide that level of service. The ultimate objective of the LOS/CE study is to provide the SEEDS Formulation Team with a capability to estimate the cost for various system architectures and mission profiles. Successful development of a life cycle cost estimation capability will be dependent on an accurate assessment of the levels of services needed from ESE data service providers.

"Levels of service" will be associated with certain functional requirements, describing different degrees of performance with which the requirement would be met. For example, a functional requirement might be: "The data service provider shall distribute data and products to users on media". Accompanying this requirement might be descriptions of quantitatively distinct levels of service, such as "delivery on media shall be provided within one working day of receipt of a data request", "delivery on media shall be provided within two calendar weeks of receipt of a data request", and "delivery on media shall be provided within one calendar month of receipt of a data request". Which level of service would be most appropriate ('recommended') or acceptable ('minimum') for a particular ESE data service provider would depend on its particular mission and the needs of its users.

The first objective of the LOS/CE study is to assist the Formulation Team in establishing the minimum (and recommended) levels of service (LOS) for ESE data service providers. These LOS will be refined in a bottoms-up manner through community workshops of potential providers and users of ESE data service providers.

The second objective of the LOS/CE study is develop a suite of costs estimation tools that will enable the Formulation Team to estimate the cost impact for various architecture trades, provide NASA Headquarters with estimates of the costs for implementing varying ESE mission profiles and implementation options, and packaging the cost estimation tool kit for use by Earth Science Enterprise scientists responding to new mission opportunities in order for them to estimate the costs for developing and operating the science data ground system for their proposed mission.

The purpose of working on these two objectives together is to ensure that the cost estimation process is tied to a reasonable requirements / levels of service set.

LOS / Cost Estimation Study Overview

This section provides an overview of the LOS/CE study. It spans the work being done by GSFC staff, Dr. Bruce Barkstrom of LaRC, as well as SGT.

LOS/CE Study Tracks

The LOS/CE study is proceeding down a set of parallel tracks, including parallel cost model development efforts. These are seen as a strength, providing an ESE planner or a PI planning a mission with two or three results, and a sense of where and why they differ, will be better grounds for planning a budget than any single estimate. The threads can be held consistent by a common base, the functional areas and at some level the levels of service, and will borrow freely from each other as each refines its description of levels of service at the level of detail appropriate to it. In other words, Dr. Barkstrom's model may operate at a finer level of detail, and so will rely on a more detailed description of levels of service than SGT's model, but the two will operate from a consistent base. Similarly, neither of the cost efforts will contradict the baseline levels of service description produced by the first track with input from and reviewed by the user community.

The LOS/CE Study tracks are:

1. Requirements / Levels of Service

This track involves the development of a baseline set of requirements / levels of service, at a level of detail sufficient to be meaningful to the user, whether a research scientist 'end user' or an intermediate provider or a mission planner or data manager.

The draft levels of service will be updated per input received at the workshop. This will require a second look at the way the information is presented, and at the level of detail presented. The goal will be for the new draft to be as easy as possible for users to evaluate - meaningful but not overly detailed, with levels of service statements with requirements implied, rather than stated explicitly with associated levels of service. User feedback on the LOS draft will be sought.

2. Cost Estimation by Analogy Model

This track involves the development by SGT (Greg Hunolt, Bud Booth) of a cost estimation model using cost estimation by analogy, tied to a set of levels of service and associated requirements spanning all areas of cost at a level of detail consistent with the data available to build the database of comparables the model will rely on, and its inputs and outputs.

Scenarios will be developed to explain the use of the cost model, with the key being to offer the user an unconstrained menu of functions / LOS choices, rather than require that the user select an a priori data service provider type to develop an estimate for. Scenarios will also be developed to explain the use of the cost model to estimate ESE enterprise level costs for alternative architectures of data service provider entities, each embodying one or more logical data service provider types. Feedback on the scenarios will be sought from potential users of the cost model.

3. Barkstrom Cost Model

This track involves a parallel effort (by Dr. Bruce Barkstrom) to produce cost estimates operating at a finer level of detail and employing deeper analytics that will produce a second set of results. Dr. Barkstrom's approach to building a cost model for evolving ESE data systems is:

a. Start with data life cycle, typically "Prepare, Validate & Produce, Use". "Validate & Produce" activities are dependent on the number of products and the number of versions of each product. Each version goes through four phases: 1) remove blunders, 2) checkout current version, 3) reduce backlog of delayed data, 4) produce data for current stream. "Validate and Produce" activities for both science team and operations staff are proportional to average number of jobs run per month, with adjustment for degree of automation.

b. Add in data use activities: use Innovation-Diffusion model for spread of data product understanding into the user community. Base staff activity on number of "sales assistance" and "troubleshooting" calls per 1000 user orders. Add in outreach and continuing evolution activities.

c. Collect data missions into organizations, and add in management and infrastructure components of services.

Underlying these three stages, there will be service choices that can be applied to each phase. For each service choice, there will be technology choices that have both an investment component and a cost component.

This kind of model lends itself to Monte Carlo or simulation approaches, which should make it possible to explore more possibilities than we would otherwise be able to do.

4. Application of COTS Cost Estimation Tools

The LOS/CE study team (David Torrealba, SGT) has completed its survey of commercial-off-the-shelf (COTS) cost estimation tools and recommended acquisition of a suite of tools for estimating lifecycle costs. This suite includes software tools for demand forecasting, neural networks, and case-based reasoning as well as traditional parametric modeling. Using these tools, the study team will investigate potential synergies with cost estimation by analogy and develop a prototype cost model for evolving ESE data systems.

The study team identified three approaches for investigating the potential synergies:

1) Non-algorithmic (‘machine learning’) techniques for estimating costs by analogy

2) Demand models based on time series analyses of user access and distribution data

3) Function point analysis of ‘dataflow centered’ models familiar to the science community

After the analogy model (see above) or the machine learning tools (neural networks and case-based reasoning) identify analogies in the ‘comparables’ database, size data (SLOC or function points) can be applied as inputs to a parametric model. The advantage of this approach is that one will then be able to use Constructive Cost Model (COCOMO) II capabilities to describe and account for differences in development environments. COCOMO II is ‘open source’ code. Its relationships, algorithms and interfaces are publicly available and well defined. Tools for cost estimation by analogy can be combined with COCOMO II in a relatively straightforward manner.

A ‘technical’ approach to modeling user demand will employ a standard forecasting tool to perform time series analyses of user access and distribution data. In this case, the investigation will focus on how user demand, if known, may apply to cost estimation models. Part of this work will be to determine the usefulness (for calibrating cost models) of currently available data.

Function point analysis, which is included as the size estimation ‘front end’ of many parametric models, may support a ‘dataflow centered’ approach that is more familiar to users in the science community. Function points are the weighted sums of factors that relate to user requirements for data management including numbers of inputs, outputs, logical files, inquiries, and interfaces.

Intended outcomes of continuing this investigation of COTS cost estimation tools are (1) to assist in test and evaluation of a cost model using cost estimation by analogy, (2) to prepare a prototype ‘COTS tool kit’ to support an analogy model, and (3) to demonstrate a cost model at the Community Workshop in June 2002.

5. Information Collection - Building the Comparables Database

A major effort has begun to collect the information from existing data activities that is needed to build the comparables database. The two step approach that has been adopted for this effort has now begun. A first round of letters has been sent to data centers (by email) asking for documentation or pointers to documentation holding the answers to a set of questions (attached to the letter). The LOS/CE team will attempt to develop the answers to the questions by researching the documentation, and only go back to the data centers for clarifications and to fill gaps.

This process will be refined in response to feedback received during its first round. The area of inquiry (and thus the question list) will be tailored to the intended recipient, and potential recipients will be researched in advance to ensure that only those most germane to the study are asked to participate.

This effort will proceed for many months. The near term intent is to get a sufficient sample to support model development and demonstration of a prototype capability as soon as possible.

6. Community Feedback

The LOS/CE study will not be successful without feedback and guidance from the (at least partially overlapping) community of users and data service providers. The February, 2002, SEEDS Workshop, while representing a first step, generated a number of recommendations about how feedback should be sought. A workshop scheduled for June, 2002, will concentrate on issues of importance to the user community, including how to increase involvement of users in this process, and presenting 'best practices' from organizations internal and external to NASA. Feedback to the model includes (but is not limited to) comments given at workshops, answers to questionnaires, providing white papers for general consumption, being a 'SEEDS prototype', or participating in this study team (for example, by 'tire kicking' model prototypes in the future). Once the model is generated, continued use will enable iterations for improved prediction capability.

Summary Schedule

February 19, 2002 - Began site information collection effort (continues for a year or more).

March 8, 2002 - Posted Vanessa's workshop results and next steps presentation to the SEEDS website. Follow with the synthesis of workshop results and next steps paper.

March and April, 2002 - Convert February workshop white paper into a set of six smaller white papers, update per workshop results (including a redo of the Levels of Service per the workshop results) and post to SEEDS website.

April, May and June, 2002 - Seek feedback on new LOS draft.

April and May 2002 - Begin building comparables database, adding Benchmark Study data and newly
collected site information as received (continues for a year or more).

May and June 2002 - Develop preliminary / placeholder cost estimating relationships for early prototype cost model.

June, 2002 - Report progress on / possibly demonstrate early prototype cost estimation by analogy model (June SEEDS Workshop).

June, 2002 - Report on COTS cost estimating tool survey, discuss most promising tools and how they can be used to support SEEDS cost estimation (June SEEDS Workshop).

September 2002 – Release prototype cost estimation tool for tire-kicking.

September 2003 – Release first version of cost estimation tool.

Technical Approach

This section outlines the technical approach being taken by SGT to its work to meet the two objectives stated in Section 3 above. The two objectives are inseparably coupled; costs must be driven by requirements, and so the cost estimation tools must be based on a model that maps directly to the requirements set. For this reason SGT's effort consists of two parallel and intertwined paths that will merge in the final product. The first path is a requirements ‘ levels of service analysis, and the second path is development of a cost estimation capability. The work on the two paths is closely coupled, as the requirements must map to the same framework as the costs, and the concept of a general data service provider reference model (see **“**Working Paper 3 - Data Service Provider Reference Model - Functional Areas”) will be used to provide the common framework.

Notes and Assumptions on the Technical Approach

This section contains notes that are background for the discussion of the technical approach to the requirements analysis and cost estimation by analogy model development that follow below in Sections 5.2 and 5.3.

Data Service Providers, the Reference Model, and Requirements

The term ‘data service provider’ is used herein as a broad, generic term for a site or activity that performs all or a subset of the functions defined in the general data service provider reference model. Many well known actual data centers such as the DAACs (Distributed Active Archive Centers) or the NOAA national data centers will perform a subset of the general list of functions, while some sites described as ‘data service providers’ for this study, e.g. MODAPS (as a sample of a SIPS (Science Investigator-led Processing System), a science team processing facility that does not perform archive or general user distribution), are different in function from many well known data centers but fit within the framework of the data service provider reference model.

The general data service reference model is defined (see Working Paper 3, “Data Service Provider (DSP) Reference Model - Functional Areas”) in terms of a set of functional areas, and a set requirements / levels of service is being developed within the functional area framework. These, documented in Working Paper 5, “Data Service Provider Reference Model - Requirements / Levels of Service”, will comprise a general set of requirements / levels of service that is independent of any physical entity or architecture. This general set is also a template, in that it contains placeholders for many specifics that would have to be defined in any real case.

The general data service provider reference model will have subsets corresponding to the tentatively defined ESE logical data service provider types (see “Working Paper 6 - ESE Logical Data Service Provider Types”, which discusses the current open set of types), seen as logical functional groupings based on an ESE role or mission as an organizing principle. This approach has the advantage of allowing the future definition of additional data service provider types, or variations of the types defined herein, i.e. other possible subsets, within the framework of the general model. In the same manner, the general set of requirements / levels of service will have subsets corresponding to each of the defined data service provider types.

Data services provider types, as logical groups of functions, do not correspond to physical entities (e.g. data centers or flight project data systems) except in a case where a physical entity performs a single ESE role or mission that corresponds to a data service provider type. In most cases, physical entities / organizations will have more complex ESE roles and missions that would correspond to some combination of the same or different logical data service provider types.

As in the case of the overall general set of requirements / levels of service, the requirements / levels of service set for a data service provider type will be also be a template containing placeholders for quantitative parameters that would be defined for a specific instance of a data service provider of that type. For example, suppose that a cost estimate is needed for a simple case where an entirely new organization is being set up to perform the functions of a single data service provider type. A requirement in the template for that type might be that "the data service provider shall provide an archive capacity of [number TB]". If the actual mission of the data service provider required that it archive certain data streams and generated products that would accumulate to a total volume of 100 TB, then that value would be inserted into the template, with the result being a specific requirement for that data service provider (i.e., "the data service provider shall provide an archive capacity of 100 TB") that could then be used in the process of generating a cost estimate for the data service provider.

COTS Cost Estimation Tools

The use of COTS cost estimation tools is being explored, for example for software development, to check the results of the cost estimation by analogy model, to provide alternate results for evaluation, or perhaps to replace it for aspects of the cost modeling where a COTS tool proves to be superior in tests against the independent cases. This requires an examination and evaluation of the available COTS tools, selecting the most promising for test, and exercising them. The SGT report "Survey of Cost Estimation Tools, Final Report" by David Torealba, February 28, 2002, reports on progress in this area.

User Model

The life cycle cost model will need to project user demand for a data service provider's services over a period of time. In addition to data service provider history information, the effort will include an examination of existing user models including Dr. Bruce Barkstrom's.

Technical Approach - Requirements Analysis

The requirements / levels of service developed by this study are intended to support the cost modeling effort, and not to serve as the complete definition of the requirements side of a contract between the SEEDS program office and ESE data service providers, or as a basis for procurements. The requirements will be ‘end-to-end’ in that they will encompass all significant elements of cost, and will be directly and explicitly traceable to cost.

The requirements analysis would proceed as follows:

a. Review existing NewDISS and ESE program documents, and incorporate the draft “NewDISS Level 0 Requirements, September 2001”, as a high level programmatic framework. Review the EOSDIS Level 2 Requirements for Version 0 as a reference to a previously defined set of requirements and levels of service that could be a source for the current effort. Review USGCRP and CES reports (see references) for additional input on requirements / levels of service.

b. Develop an initial general requirements / levels of service template corresponding to an initial version of a general data service provider reference model, and subsets corresponding to an initial set of logical data service provider types, consistent with program documents, for review, revision as needed, and approval as a starting point by the SEEDS Formulation team.

c. Use a community workshop to get user feedback on, and input into, the requirements and levels of service definitions. Produce updates to the requirements templates set, for review, revision as needed, and approval by the SEEDS Formulation team.

d. Work with the user and provider community and the SEEDS Formulation Team to obtain feedback on and guidance for the improvement of the requirements / levels of service.

e. Produce a final requirements templates set for review, revision as needed, and approval by the SEEDS Formulation team, and produce a final report on the requirements analysis. (SGT has a contract deliverable for a final set of requirements / levels of service on March 31, 2002, but will maintain and update the document as work progresses over the life of the task.)

At each step, as changes to the requirements sets are approved by the Formulation team, ensure that the requirements changes are reflected back into the data service provider reference model.

Technical Approach - Cost Estimation by Analogy Model Development

The cost estimation model will be based on a ‘comparables’ or ‘cost by analogy’ method; it will estimate costs using cost estimating relationships derived from a number of existing data service providers that are functionally comparable to the different types or combinations of different types of ESE data service providers.

The general approach is to draw on the data service provider reference model concept developed for the Best Practices / Benchmark study and to develop a cost estimation model that estimates the cost of an ESE data service provider based on the actual costs of comparable data service providers, a “cost estimation by analogy” methodology. Information about other ESE or outside data service providers will be collected to provide the best possible basis for comparison. The cost estimating relationships that will be used by the model will themselves be developed and evolve as the comparables database is built (see “Working Paper 2 - Cost Estimation by Analogy Model” for a description of how this process is seen).

The life cycle cost model will project user demand for a data service provider's services over a period of time. In addition to data service provider history information, the effort will include an examination of existing user models including Bruce Barkstrom's.

The cost estimation by analogy model development will proceed with the following steps:

a. Define, and refine based on feedback from ESDIS and the Formulation Team, the content of a data service provider cost estimate; i.e. what elements of cost at what level of detail with what supporting information are required as the output product from the cost estimation tool. The further development of the cost model would be guided by the results that the model must produce, allowing for the fact that this will change as the effort proceeds.

b. Survey available COTS cost estimation tools, evaluate and test the tools that seem most likely to be useful for this study, and produce a report summarizing the results of the survey and recommending tool(s) to be used further in the study.

c. Obtain and examine the Bruce Barkstrom user model and any other user model that might be useful for this study.

d. Extend the existing Best Practices / Benchmark study reference model to encompass the full range of data service provider functions, refine the original list of model parameters, add implementation, parameters necessary for estimation of cost, etc. (See “Working Paper 4 - Data Service Provider Reference Model - Model Parameters” for definition of the model’s parameter set.)

e. Derive subsets of the general reference model to correspond to the logical data service provider types. These subsets will include the functional areas and metrics appropriate to each data service provider type.

f. Map the information collected on selected data service providers during the Best Practices / Benchmark study to the extended reference model to begin to build the model's comparables database;

g. Identify additional data service providers to be added to the comparables database. Draw from DAACs not included in the Best Practices / Benchmark study, SIPSs, selected ESIPS. Consult with ESDIS to arrive at a list of candidates. Reserve some data service provider cases for use as independent test cases for the cost model.

h. Collect the additional or update information and add to the model's information set (i.e., as was done for the Best Practices / Benchmark study, map data service provider information to the reference model's common set of metrics).

Note that steps b and c can run in parallel with a, d, etc. Also, steps g and h can run in parallel with d, e, and f provided that information collected early on can be supplemented as completion of steps d, e, and f identify gaps in the initial collection.

i. Use the mapped data service provider information to construct relationships (for each data service provider type, within each functional area) relating actual data service provider staffing and costs and known development effort and workload performed, etc. These relationships are currently TBD but could include linear regression equations and the like. Probable errors of estimate will also be derived for each relationship.

j. Test the model by inputting information for the independent test cases and determining the degree to which model is able to correctly calculate staffing, costs, etc. Test the COTS cost estimation tools, to determine which should be incorporated into the model or used in conjunction with the model to give the best possible overall result. Also consider incorporation of the Bruce Barkstrom or other externally developed user model.

k. Obtain community feedback on the prototype cost estimation model.

l. Release a life-cycle cost model Version 0 that incorporates initial user feedback; continue to obtain and incorporate community feedback by presenting study results and providing prototype models for hands-on peer review.

m. Provide a final report, and provide cost estimates for ESE data service providers as needed.

SEEDS

Levels of Service / Cost Estimation Study

Working Paper Two:

Cost Estimation by Analogy Model

May 15, 2002

G. Hunolt, SGT, Inc.

Outline

Introduction

This working paper is the second of a set of papers that describes the SEEDS (Strategic Evolution of Earth Science Enterprise Data Systems) Levels of Service (LOS) and Cost Estimation (LOS/CE) study. The study goal is to develop a cost estimation model and coupled requirements and levels of services to support the SEEDS Formulation team in estimating the life cycle costs of future Earth Science Enterprise (ESE) data service providers and supporting systems, where 'data service provider' is used as a generic term for any data/information related activity. The set of working papers is intended to serve as a vehicle for coordinating work on the project, obtaining feedback and guidance from ESDIS (Earth Science Data and Information System project) SOO (Science Operations Office) and the user community, and as embryos of reports that will be produced as the task proceeds.

As working papers, each version of each working paper that appears represents a snapshot in time, with the work in various stages of completion; readers should expect loose ends and inconsistencies especially in the early stages of the project. As work progresses the content (and sometimes the organization) of the working papers will change reflecting progress made, responses to feedback and guidance received, etc.

Introduction to Working Paper 2 - Cost Estimation by Analogy Model

This second working paper of the set describes the Cost Estimation Tool and the underlying cost estimation by analogy model that is being developed by SGT for this study. (The Cost Estimation Tool is simply the packaging of the model in a usable form, including a user interface and report generating capability.) This paper will evolve extensively as the work progresses. Its initial focus will be on a conceptual description of the tool and the model, and how the model and the cost estimating relationships it will use are expected to develop, requirements and operations concepts including scenarios showing how the tool will be used, goals and plans for the demonstration prototype. A major update to this working paper will be provided by June 30, 2002 that will address plans for development beyond the demonstration prototype and for test and evaluation of the prototypes, including "tire-kicking" by users.

As a part of the LOS/CE Study, and in parallel with the effort described in this working paper, SGT is examining COTS cost estimating tools (e.g. parametric cost models) to see if one or more of these might be better for certain aspects of costing than the cost model to be developed during the study, or valuable for use in producing alternative cost estimates for some or all aspects of costing. At least some COTS tools can be integrated with other software such as Excel, and so it may be possible to deliver a cost estimation tool with an integrated COTS component. In any case, the most practical approach will be taken to facilitating the use of any selected COTS tool in conjunction with the model developed by the study.

As the needs of the ESE science and applications program evolve, and hence the ESE roles and missions for data service providers evolve, and as information technology that touches all aspects of every data service provider and the user community evolves (e.g. the Grid distributed computing approach), the data service provider reference model and the cost model will evolve. The content of this paper can only represent a snapshot in time - and indeed a snapshot that is in part tied to current and recent past experience with working data service providers. If the cost estimation tool (and the underling data service provider model) proves useful, it will have to be maintained and revised perhaps dramatically to preserve or improve its usefulness over time.

Technological changes and the evolving requirements of science and applications users, perhaps especially in the access and distribution area, call into question the ability of the cost estimation tool to make reliable cost estimates for the future based on current and recent past experience. The problem is acknowledged, and as the tool is developed and tested and its sources of error are analyzed, this aspect will not be ignored.

In addition to evolving with changing ESE program needs, the cost estimation by analogy model (and the data service provider model) will be improved in successive iterations as the comparables database grows and includes more new activities, and with lessons learned derived from use of earlier versions of the model.

Section 2 discusses the objective, scenarios, and requirements for the Cost Estimation Tool. Section 3 describes the output to be provided by the Cost Estimation Tool. Section 4 discusses development of the cost estimation by analogy model that is the heart of the Cost Estimation Tool, and discusses the first phase of cost estimation tool development, the demonstration prototype.

Definition of Key Terms

This section defines a few key terms that will be used frequently in the remainder of this working paper.

As noted above, the term ‘data service provider’ is used as a generic term for any data/information related activity, such as a data center (e.g. the EOSDIS Distributed Active Archive Centers (DAACs), a flight project data system (e.g. the MODIS Adaptive Processing System (MODAPS) or TRMM Science Data and Information System (TSDIS). A data service provider provides all or some subset of the functions described by the general data service provider reference model (see Working Paper 3, “Data Service Provider Reference Model - Functional Areas”), which include data ingest, processing, archive, distribution, etc. The scale and scope at which these services are provided depends on the particular ESE role or mission responsibility of the data service provider.

A data service provider is an element of, or is operated or hosted by, an ESE or ESE funded organization, which might be dedicated to supporting a single data service provider, a number of data service providers, or a combination of a data service provider and other different activities.

The term ‘data service provider’ is also used in an abstract sense in the context of logical data service provider types, as described in Working Paper 6, “ESE Logical Data Service Provider Types”, which are essentially functional groupings, subsets of the general data service provider reference model, organized around an ESE general role or mission. Logical data service provider types do not necessarily map one to one to physical entities, i.e. to actual operating data service providers. A real world physical data service provider, such as a DAAC, often will embody more than one of the logical data service provider types, depending on the complexity of its specific ESE role or mission.

The term ‘data service provider activity’ refers to the work performed by a data service provider to meet the needs of a particular project or research or applications effort, regardless of whether the data services provider is an integral part of the project (such as a flight project data system like TSDIS or MODAPS) or whether the activity is performed by an organizationally or physically separate data services provider, or a hybrid split between those two cases (such as when a flight project does its own product generation but ‘subcontracts’ archive and distribution to a DAAC). A working data services provider may engage in a single activity, or multiple activities if it supports multiple flight projects or research / applications efforts, perhaps in addition to a core ESE data management mission.

In the most complicated case, a single ESE organizational element or ESE funded organization may serve as an ESE data service provider, embodying a number of logical data service provider types, and within each type, perform a number of activities supporting a number of projects or research / application efforts.

Cost Estimation Tool Scenarios and Requirements

This section discusses use case scenarios and requirements for the Cost Estimation Tool, where the term ‘Cost Estimation Tool’ is used to mean the cost estimation by analogy model packaged in a useable form, i.e. provided in a package that can be started up, can receive a set of inputs, run, and produce a set of outputs. An objective of the study is to provide the tool in as readily useable form as possible, for example as an Excel spreadsheet workbook that could loaded and used on any PC or Macintosh platform equipped with Excel.

As noted above, the Cost Estimation Tool is needed to enable the SEEDS Formulation Team to estimate the cost impact for various architecture trades, and to provide NASA Headquarters with estimates of the costs for implementing varying ESE mission profiles and implementation options. The Formulation Team also requires that the tool be packaged so that it can be provided to ESE scientists for their use in estimating the costs for developing and operating the science data ground system for their proposed mission.

The remainder of this section examines the particulars of this objective and what the Cost Estimation Tool must be able to do to meet it. Section 2.1 presents the use case scenarios, and Section 2.2 presents general requirements for the Cost Estimation Tool.

Cost Estimation Tool Scenarios

This section describes scenarios that describe how the Cost Estimation Tool would be used.

Several categories of use are envisioned:

1) The first is use by a flight project or science team to estimate the costs of implementing and operating a set of data management functions required to support their project or research effort. (This would constitute a single data service provider activity.)

2) The second is by an existing data center that has been asked to estimate the costs of adding an additional set of data management functions, perhaps to meet the needs of a new flight project or research effort. (In this case the data center would be adding an additional data service provider activity to those which it already performs.)

3) The third is by an ESE program office wishing to make overall estimates of implementation and operating costs for a constellation of ESE data service providers operated by a number of ESE or ESE funded organizations, collectively performing all of the data service provider activities required by the ESE program (i.e., supporting all ESE flight projects, research / applications efforts, and general data management needs). The ESE program office may wish to examine ‘architecture trades’ - alternative mappings of

functions and mission responsibilities to ESE organizations, perhaps to make long term budget estimates.

Each of these categories of use will be discussed below.

New Project or Research / Applications Effort

Assume that a new flight project or research effort is being proposed in response to a NASA Announcement of Opportunity (AO) or other solicitation vehicle. The group developing the proposal examines its need for data management support - i.e. decides what functions it requires and what service needs it has for each (i.e. what levels of service it needs). It also puts together a description of its mission requirements for data management support - e.g. best quantitative estimates of what data will be received, produced, distributed (details sensor and ancillary data streams, products to be generated, distribution to team members or other users) etc.

The group is now ready to use the Cost Estimation Tool, and it is assumed that one group member proceeds as follows:

a. The user activates the tool, using a stand-alone distributed version or a web accessible version.

b. The user selects from a menu of functions those that are needed to meet the project's needs, and, for each function:

1) The user selects levels of service to be provided, as applicable or needed for the particular function.

2) The user provides quantitative mission detail as applicable for the particular function.

3) The user provides re-use factors, to account re-use of existing capabilities, if any.

c. The user provides control parameters such as implementation start date, operations start date, projected activity end date, etc.

d. The user provides costing parameters such as inflation rate, labor rates.

e. The user runs the model to produce the life cycle cost estimate for the data service provider activity.

The cost estimate will be a life cycle cost including year by year development and sustaining engineering costs and operations staffing and costs projected over a number of years (see Section 3 for a description of the output of the cost model).

This scenario makes no use of the predefined set of logical data service provider types, allowing the user complete flexibility in selecting needed data service functions. But the user would have the option of judging that a particular type of data service provider was a good fit, and at step b in the scenario bringing up a template for that type, and then providing the information as 'filling in the blanks' in the data service provider template.

ESE Data Center Takes on a New Task

Assume that an ESE data center, already engaged in a number of data management activities, perhaps supporting one or more flight projects or research efforts, wishes to propose to perform an additional data management task in response to a NASA AO, or has been asked by another group preparing a response to a NASA AO for (for example) a flight project to propose to provide data management support. Guided by the functions that would be required of it (in the examples given, either by the AO or the group preparing the flight project response), the data center would assemble the description of the new task requirements for data management support - e.g. best quantitative estimates of what data will be received, produced, distributed (details sensor and ancillary data streams, products to be generated, distribution to team members or other users), etc. The data center will also determine what ability it will have to "reuse" its existing infrastructure to support the new task (staff, systems, facility, etc.).

The data center is now ready to use the Cost Estimation Tool, and it is assumed that one data center staff member proceeds as follows:

a. The user activates the tool, using a stand-alone distributed version or a web accessible version.

b. The user selects from a menu of functions those that are required to perform the new task, and, for each function:

1) The user selects levels of service to be provided, as applicable or needed for the particular function.

2) The user provides quantitative mission detail as applicable for the particular function.

3) The user provides re-use factors, to account the data center's re-use of existing capabilities, if any.

c. The user provides control parameters such as implementation start date, operations start date, projected activity end date, etc., for the task.

d. The user provides costing parameters such as inflation rate, labor rates.

e. The user runs the model to produce the life cycle cost estimate for the data service provider activity.

The cost estimate will be a life cycle cost including year by year development and sustaining engineering costs and operations staffing and costs projected over a number of years (see Section 3 for a description of the output of the cost model).

This scenario makes no use of the predefined set of logical data service provider types, allowing the data center complete flexibility in selecting needed data service functions according to the requirements given it by the AO or the flight project (for the two possible cases noted). But the user would have the option of judging that the data center was a good fit for a particular type of data service provider, and at step b of the scenario bringing up a template for that type, and then providing the information as ‘filling in the blanks’ in the data service provider template, in that way ‘tuning’ the template for the data services provider type to meet the given requirements.

ESE Architecture Trade Study

This section describes the use of the Cost Estimation Tool to support ESE architecture trade studies.

In this context, the term ‘ESE architecture’ means a collection of physical entities, i.e. ESE or ESE-funded organizations, performing a set of data service provider activities that in their aggregate meet ESE’s requirements for data management and services support to its flight projects and research and applications programs.

An architecture would be developed by analyzing the complete set of ESE program requirements for missions to be supported, science and applications efforts to be supported, data holdings to be maintained, the data service needs of various user communities within and without the ESE program, etc., and determining a set of ESE data service providers (and their required interconnections) able to meet the various ESE mission requirements. There will be many possible architectures, i.e. many possible configurations of ESE data service providers, that will meet a given set of ESE mission requirements, and hence the need for the ability to analyze trades between them to select the architecture to implement.

ESE mission requirements and hence the needed set of ESE data service providers will vary over time, given factors such as the launch dates for flight missions, the phasing of efforts in the science program, possibly the rotation of data into non-NASA long term archives, etc. As a result there will not only be many possible architectures at any given time, but also many possible paths for the evolution of the overall ESE architecture as time goes on. That is to say there will be many, time varying configurations of ESE data service providers that comprise possible ESE architectures. These may represent different approaches to consolidation of new mission requirements at ESE data service providers, different assignments of level of service requirements, different consolidations of new mission requirements with ongoing (i.e. EOSDIS) mission requirements, etc., and combinations of the above.

For each possible architecture, the cost estimation tool would generate cost estimates for the individual components and allow a roll-up of these to accumulate to an overall estimated cost for the candidate architecture. This would then be one factor taken into account in consideration of the trades between possible architectures (along with complexity, robustness, etc.).

Assume that an ESE data services architect wishes to generate a cost estimate for one of the many possible architectures. The first step is for the architecture to be defined as a mapping of data service provider activities to existing or new (presumably placeholders for winners of future competitions) ESE organizations. To simplify this task, the architect can use the defined logical data service provider types, and think of the problem as assigning data service provider activities or mission responsibilities (e.g. for supporting a flight project, supporting a research effort, providing general access to a large data collection, etc.) to a set of instances of data service provider types, and then mapping the instances of data service provider types to existing and/or new ESE organizations.

The ESE data services architect must also have specific mission requirements for the various flight projects, science efforts, etc., that comprise the ESE program. While an individual project (as in the cases described above) could be expected to know / forecast its own requirements in some detail, the ESE data services architect is likely to have to make rough estimates. This would reduce the accuracy of the result for each individual component of the architecture, and the error of estimate of the ESE architecture as a whole would be greater than that associated with an individual estimate produced by a flight project or data center as described in the previous sections.

Now the ESE program office user has a set of logical data service providers defined, and a set of organizations to which these are mapped, and mission requirements for flight projects, science / applications efforts, etc. The user could use the Cost Estimation Tool as follows:

a. The user could first use the Cost Estimation Tool to obtain an estimate for each logical data system provider as if it were to be physically a stand-alone entity performing a specific data service provider activity - i.e., mapped to one ESE organization dedicated solely to it, following the scenario in Section 2.1.1 above. The sum of all these would provide an aggregate ESE level cost estimate that would represent a worst case - having a separate physical ESE organization for each logical data service provider activity would result in no savings from reuse of infrastructure.

b. The user would then apply the mapping, i.e. the architecture:

1) For each ESE organization that as an element of the architecture, the user would select one of its logical data service provider functions as a base, and produce an estimate for it as described in Section 2.1.1.

2) Then the user would make separate estimates for the incremental costs of adding each of the other data service provider activities following the scenario described in section 2.1.2, until the total for that organization was reached. This would include taking re-use into account.

3) The user would repeat the process for all of the ESE organizations comprising the architecture.

c. The user would sum up the estimates for the individual organizations and arrive at an estimated cost for the specified architecture, including an overall error of estimate.

d. The user could then modify the architecture, and redo the process, to obtain another estimate for comparison.

Cost Estimation Tool Requirements

This section outlines the basic requirements the Cost Estimation Tool (a.k.a. the Cost Tool) shall be capable of meeting, following the scenarios in Section 2.1. The requirements are divided into two groups, functional and implementation.

Functional Requirements

The Cost Tool shall be based on a cost estimation by analogy approach, supplemented by parametric techniques as needed or advantageous.

The Cost Tool shall estimate full life cycle costs for a data service provider activity, from the beginning of implementation through a specified operational life time.

The Cost Tool shall estimate costs for all areas of cost, including implementation, operations, maintenance and management staff, COTS hardware and software, custom software, logistics and supplies, etc. See Section 3 for a detailed description of the required Cost Tool output.

The Cost Tool shall provide errors of estimate for each projected cost.

The Cost Tool shall allow the user to select functions to be included in the estimate for a data service provider activity and provide parameters for each function appropriate to the data service provider activity's mission or ESE role.

The Cost Tool shall, as an alternative mode of operation, allow a user to select from a set of logical data service provider types, each including a pre-selected set of functions.

The Cost Tool shall support the estimation of costs for a data service provider activity as either a new, stand-alone activity (such as a new data system to be implemented as part of a NASA flight project) or as an incremental increase to an existing data service provider activity (such as the incremental costs to an existing NASA data center to provide archive and distribution support to a NASA flight project), in which case the Cost Tool user shall provide appropriate re-use factors.

The Cost Tool shall support subcontracting or partnering, such as allowing for a new NASA flight project arranging for an existing NASA data center to provide archive and distribution support.

The Cost Tool shall support generation of cost estimates for an ESE data services architecture, including overall error of estimate.

Implementation Requirements

This section describes implementation requirements for the Cost Estimation Tool.

The Cost Estimation Tool shall be capable of stand-alone operation on any reasonably sized PC or MacIntosh platform, as an Excel workbook application or equivalent.

The Cost Estimation Tool shall be fully documented, including a users' guide.

The precise form of the cost estimation tool is TBD as of now, but might include an integrated COTS tool, or the package might include both an Excel based tool and a separate COTS tool or tools. In any case, the package will be documented as a whole, with a users' guide covering the entire package, supplemented by COTS documentation as needed.

ESE Data Service Provider Cost Estimate Content

This section describes the content of cost estimates to be provided by the Cost Estimation Tool, i.e. a statement of what the cost estimate provided by the tool will consist of. This description of the model output will drive requirements for the combination of input and computation needed to produce it.

The Cost Estimation Tool must provide estimates of implementation and operating costs, over a specified lifetime, for a data service provider activity (as a stand-alone or increment). The implementation period costs must include hardware purchase, custom software development and COTS purchase, integration and test costs, and facility preparation costs. The operating period costs of a data service provider activity must include hardware maintenance, continuing COTS support, sustaining engineering, operations, recurring facility costs, supplies such as storage and distribution media, and must allow for the possibility of 'technology refresh'. Implementation and operations period staff costs must allow for reasonable management staffing, and labor rates must allow for overhead and inflation adjustments.

Tables 1, 2, and 3 below are together an initial example of what the cost estimation output might look like. The categories would be defined in detail below. Note that the actual number of years for which costs would be estimated (shown as seven in the example) would be selectable as appropriate for actual cases.

Table 1 - Draft Sample of Cost Estimation Output - Implementation Period Costs

	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6	Year 7	Total
Estimated Implementation Costs								
Management Staff, FTE								
Management Staff Cost								
Development Staff, FTE								
Development Staff Cost								
Hardware Purchase								
COTS Software Purchase / License								
Facility Preparation								
Total Implementation FTE								
Total Implementation Cost								

The cost estimate example shown above and below contains some FTE lines that would be generated by the model in the process of producing the cost estimate. Other such lines could be added, such as SLOC to be developed and maintained.

There are a variety of “workload” parameters that could be presented in conjunction with the cost estimate. These could include those that characterize the mission of the ESE data service provider, which would have been provided as input to the cost estimation, such as flight mission to be supported, input data streams, output product streams, etc., appropriate to the type and particular mission of the data service provider.

Table 2 - Draft Sample of Cost Estimation Output - Operations Costs

	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6	Year 7	Total
Estimated Operations Costs								
Management Staff FTE								
Management Staff Cost								
Technical Coordination Staff FTE								
Technical Coordination Staff Cost								
Sustaining Engineering FTE								
Sustaining Engineering Cost								
Engineering Support FTE								
Engineering Support Cost								
Operations Staff FTE								
Operations Staff Cost								
Development FTE								
Development Staff Cost								
Recurring Network / Comm Cost								
Recurring COTS S/W Cost								
Hardware Purchase Cost								
Hardware Maintenance Cost								
Supplies Cost								
Recurring Facility Cost								
Total Operations FTE								
Total Operations Cost								

Table 3 - Draft Sample of Cost Estimation Output - Total FTE and Costs

	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6	Year 7	Total
Estimated Total FTE								
Estimated Total Cost								

All of the parameters shown in the tables above are defined in Working Paper 4 “General DSP Reference Model - Model Parameters”, as are also the required input parameters and all of the internal parameters used by the cost estimation by analogy model.

Development of the Cost Estimation by Analogy Model

This section discusses the development of the cost estimation by analogy model that is the heart of the Cost Estimation Tool.

Cost Estimation By Analogy

The cost estimation by analogy technique is based on the idea that reasonably reliable estimates for the cost of a future data activity (either by a new organization formed for that purpose or as an increment to the data activities of an existing organization, or some combination thereof) can be based on an analysis of the past history and experience with other similar data activities.

Contained in this simple statement are some assumptions that must be taken fully into account:

1) That a sufficient sample of reasonably applicable cases exists on which to base an estimate.

2) That cases are either applicable in implementation and operation approach as well as function and workload, so that the effort required for the cases can be taken as suggestive of the effort that will be required for a new case to be estimated.

The first assumption is important when statistical techniques such as regression are considered; if there is too small a sample the results will be unreliable or entirely useless, as will be indicated by the probable errors of estimate that will accompany the estimates, and by the results of tests on independent cases.

The second assumption reflects the concern that a project that might be nearly identical in terms of the nature of the data activity (function and workload) to be estimated but might have been done (implemented and/or operated) by an approach so different as to compromise partly or completely its value as a data point for producing an estimate for a new activity. Attention must be paid to trends that could follow changes in approach that might provide a basis for an extrapolation into the future.

An important point that must be made at the outset is that the model will not be estimating future costs on the basis of past costs. It is indeed almost a misnomer to call the model a ‘cost model’ because the real basis for comparison with cases is staff effort and system capabilities. Year by year costs are only added as a final step. A year by year effort estimate is first produced, and then priced out by application of labor rates and inflation. Similarly projections of required system capabilities are made, and then priced out through use of system capability vs projected cost curves. Other non-staff elements of cost are handled in like manner. Finally all factors are summed to produce the final output, the year by year life cycle cost estimate.

General Development Considerations

This section outlines some general considerations regarding the approach being taken to developing the cost estimation by analogy model.

The model will evolve over the life of the project. This evolution will be driven by a number of forces: 1) the building of the comparables database, 2) the feedback obtained from users evaluating prototypes of the Cost Estimating Tool, 3) the feedback gained from experience with actual use of the Cost Estimating Tool, and 4) the progressive growth of the comparables database, as new cases are added and as the information about existing cases is updated.

The building of the comparables database will drive the evolution of the model. This is because the cost estimating relationships used by the model are dependent on, or constrained by, the state of the comparables database. The types of cost estimating relationships to be used by the model are described in the next section. Those relationships which draw on the comparables database - the true estimation by analogy relationships - have to be developed through analysis of the available data. The state of the available data will develop slowly as the information collection process goes on - i.e. as the comparables database is gradually built. In the case of some parameters, a sufficient number of comparable cases will be accumulated to enable statistical relationships to be used. For other parameters this will not be the case, and either reasonable arithmetic approximations will be used or the parameters will have to be dropped. Thus the model has to be flexible to accommodate changes to the cost estimating relationships as more is learned about the data that will be available and various possible combinations of parameters are tested to see which combinations yield the strongest relationships.

At first only simple relationships will be employed, but as development proceeds the use of non-linear relationships will be explored, and perhaps tools / techniques that evaluate the relative ‘distance’ of the input case to the members of the set of comparables to produce a better estimate.

The general approach to the development will be to work top down, to try to come up first with the simplest possible set of cost estimating relationships, even dummy placeholders, but a working model that demonstrates how the model will run, the user interface, the output to be produced.

If the model is thought of as a Fortran program - the development approach is to have a working main program with input and output subroutines routines and dummy computational subroutines that runs so users can see and provide feedback on how it works. Imagine that each arithmetical relationship or comparables-based CER is a subroutine. The model will begin with an initial set of subroutines, each with an output and set of inputs, that uses a very simple method of doing its computation. Then, as time goes on, each subroutine will be replaced with more advanced versions which produce the same outputs but use better computational approaches, maybe a linear regression type relationship that someday might itself be superseded by a different one based on a better set of inputs or a non-linear relationship. The intent is to have a "working" model right from the start, that will gradually get better as its parts are improved.

Cost Estimating Relationships and Model Parameters

The output of the Cost Estimation by Analogy Model (or the tool which embodies it) is a list of parameters (see Section 3 above). Each of these output parameters is related to (computed from) combinations of other parameters in several steps, ultimately tracing back to either input parameters provided by the model's user or to parameters obtained from the comparables database, or combinations of the two. The term 'chain' will be used for the full sequence from an output parameter back to its ultimate inputs. Each chain consists of one or more 'links'; i.e. each chain from an output parameter back to the user input and/or comparables input from which it is ultimately derived can be thought of as one or more steps, each a link in the chain. Each link consists of an [output][process][input] sequence, with the links being connected by an overlaps between inputs and outputs. For example if output parameter A was computed from intermediate parameter B which in turn was computed from an input parameter C, then the chain would have two links, [A][process][B] and [B][process][C], with the connection between them being the parameter B which is the input to the second step and the output from the first step.

The [process] portion of each link contains a rule for computing the output from the input, known generically as a "cost estimating relationship" regardless of whether or not the output is a cost per se. The model employs three kinds of cost estimating relationships (CERs) - 'plug value', 'arithmetic', and 'comparables based'. Each of these kinds is described below.

Detailed documentation of the chains and links showing the relationships between the parameters and describing the computational steps will evolve as the project progresses. Copies are available by request.

CERs of the First Kind - 'Plug Values'

The first kind of CER is the 'plug value'. Plug values are constants plugged into the value when there is no better way of computing the output parameter. For example, a parameter may be defined to capture the level of effort required for participation in the process of developing and maintaining standards. A level of 0.25 FTE might be assessed as reasonable for this, in the absence of good documentation of actual levels of effort in past situations, or any basis for computing a level. That value, 0.25 FTE, becomes a plug value for the parameter.

CERs of the Second Kind - 'Arithmetic'

The second kind of CER is 'arithmetic'. In this case there is a simple arithmetic relationship between the output and its input(s). For example, suppose that an output parameter is the cost for management staff, and the inputs are effort in FTE, management labor rate, and the inflation rate. The arithmetic process to get the cost would be to multiply the FTE by the labor rate and apply the inflation adjustment to the result.

CERs of the Third Kind - 'Comparables Based'

The third kind of CER is 'statistical'. The output parameter is computed by a relationship that involves the data (i.e. one or more parameters) from the comparables database. The relationship may be based on linear regression, a non-linear relationship, or some more complex technique. An error of estimate will accompany the result.

The information in the comparables database (though assembled on a site by site basis) will be used on a parameter by parameter basis within the reference model's functional areas. The 'best fits' for a projected new data activity's ingest area might includes cases that were not good fits for other areas, etc.

Demonstration Prototype Cost Model - Development Approach

This section described the approach being taken to development of the demonstration prototype of the Cost Estimation Tool and the cost estimation by analogy model it contains.

Objectives for the Demonstration Prototype

These are the minimum objectives for the demonstration prototype.

1. The demonstration prototype will show how the Cost Estimation Tool will work, how a user will use it, how the scenarios in Section 2 above will be realized. The prototype will show a user picking from a general function list - i.e. be based on the general reference model and not show the data service provider subsets. The model has to do a complete execution, regardless of what simplifications are necessary at this point.

2. The demonstration prototype will use a partial, very limited comparables database based on the benchmark study data plus whatever else can be collected from about a half dozen sites.

Not Objectives for the Demonstration Prototype

These are things the first, demonstration prototype will not be capable of, presented in order to be clear about expectations.

1. The demonstration prototype will not produce useful results. The ability to produce useful results depends on the database of comparables being as large as possible, allowing the best CERs, and in the prototype timeframe the information collection and building of the comparables database will have just begun. Results will not be tested against independent cases - that will come later when more data is collected and some cases can be held aside for such testing.

2. The demonstration prototype will not show how the model can be used to estimate the costs of various possible SEEDS architectures - combinations of sites that are each combinations of one or more data service providers performing data activities.

3. The demonstration prototype will not show ‘subcontracting’ or ‘teaming’ - that will come later.

Steps to Develop the Demonstration Prototype

1. Develop an initial version of the chains and their links, the relationships between the output, internal, and input parameters, with placeholders for the CERs that will be needed.

2. Develop the schema for the comparables database and a site ‘template’ (or ‘fill in the blanks’ form), populate from the benchmark data and site data as possible (starts and goes on in parallel for at least a year).

3. Create a User Interface description, input and output, i.e. what the user will see. Reference the scenarios in Section 2 above.

4. Develop an initial set of CERs.

As far as the inputs go, as the comparables database is built it will be seen which of the possible inputs shown in the parameter matrix can actually be collected or collected in sufficient number to be used. The demonstration prototype will have to make do with whatever is available.

As far as the ‘comparables based’ computation by the prototype goes, given the limited data set that will be available, it is expected that only linear relationships will be used, e.g. a simple linear regression technique with error of estimate, an Excel function. Thus the demonstration prototype will have a set of linear equations for comparables-based CERs. There will doubtless be a number, perhaps a large number of cases where there will not be sufficient data to use even simple regression. In these cases, as placeholders, arithmetic relationships based on a (documented) assumption or two.

Other CERs will be ‘plug value’ or ‘arithmetic’.

Development Beyond the First Prototype

This section will be developed between the first version of the working paper and June 30, 2002. It will address the working prototype and operational versions of the Cost Estimation Tool. The basic approach will be a progressive refinement of the model and its CERs as the comparables database grows and as results of testing are taken into account. “Tirekicking” of prototypes of the Cost Estimation Tool by users will be undertaken and is expected to provide valuable feedback.

References and Acronyms

The References Section and the Acronym List for all of these Working Papers is in the document "References and Acronyms for the Levels of Service / Cost Estimation Working Papers ".

SEEDS

Working Paper Three: Data Service Provider Model, Functional Areas

April 24, 2002

G. Hunolt, SGT, Inc.

Outline

Introduction

This working paper is the third of a set of papers that describes the SEEDS (Strategic Evolution of Earth Science Enterprise Data Systems) Levels of Service (LOS) / Cost Estimation (LOS/CE) study. The study goal is to develop a cost estimation model and coupled requirements and levels of services to support the SEEDS Formulation team in estimating the life cycle costs of future ESE data service providers and supporting systems, where ‘data service provider’ is used as a generic term for any data/information related activity. The set of working papers is intended to serve as a vehicle for coordinating work on the project, obtaining feedback and guidance from ESDIS (Earth Science Data and Information System project) SOO (Science Operations Office) and the user community, and as embryos of reports that will be produced as the task proceeds.

As working papers, each version of each paper that appears represents a snapshot in time, with the work in various stages of completion; as work progresses the content (and sometimes the organization) of the working papers will change reflecting progress made, responses to feedback and guidance received, etc.

This third working paper of the set describes the general data service provider reference model developed for the LOS/CE study, and discusses the functional areas included in the model.

The functional area descriptions in this paper reflect results of the February, 2002, SEEDS Community Workshop, comments and recommendations made at the workshop and in white papers submitted to the workshop.

Data Service Provider Reference Model

This section describes the Data Service Provider (DSP) Reference Model, a functional model of a generic data service provider.

The reference model has three related aspects:

1) A set of 'functional areas' that collectively comprise the full range of functions that a data service provider might perform and the areas of cost that must be considered by the cost estimation by analogy model.

2) A set of parameters for each functional area that constitute a quantitative description of the workload, staff effort, and any other factors that contribute to cost for that area, additional 'roll-up' parameters that sum items such as staff effort across the functional areas, and other parameters like labor rates that are required for cost estimation.

3) A set of requirements and levels of service for each functional area.

These three aspects of the model are closely coupled to ensure the internal consistency of the model. The set of functional areas is the underpinning; both the model parameters and requirements / levels of service are organized according to the functional areas. The requirements / levels of service and the model parameters are coupled in that the definitions of the requirements / levels of service embody model parameters. This integration of the three aspects of the model is intended to ensure that estimated costs are driven by and traceable to requirements to the fullest extent possible.

The intent of the descriptions of the functional areas (see Section 3 below) and the corresponding requirements / levels of service (see Working Paper 5, "General Data Service Provider Reference Model - Requirements / Levels of Service") is to provide a reasonably full description of the abstract ESE data service provider, and to reflect the concerns expressed in the February, 2002, SEEDS community workshop. The ability of the cost estimation by analogy approach to reflect the full range of detail described in the functional areas and requirements / levels of service will be limited by the information available in the comparables database and the feasibility of reasonable assumptions where information is not available. This will be reflected in the reference model's parameter set.

Functional Areas - Areas of Cost

The functional areas of the reference model are defined in Section 3 of this paper. Some of the areas are not strictly speaking "functional" in nature (such as 'facility / infrastructure') but are needed to ensure that all significant cost areas are included.

The functions / areas of cost span the full life cycle from implementation through operations. Implementation includes capital and staff costs associated with developing, implementing, integrating and testing the data service provider's data and information system, and facility start-up / preparation costs. Implementation is assumed to be spread over a specified number of years. Implementation can overlap the start of operations. Implementation can also recur during the operating period, allowing for system expansion, enhancement, or replacement, i.e. 'technology refresh'. Operations includes hardware maintenance, sustaining engineering, operations staff, supplies (e.g. storage and archive media), recurring facility costs, etc., for the expected lifetime of the activity.

Reference Model Parameters

The parameters of the reference model are defined in detail in Working Paper 4, "General Data Service Provider Reference Model - Model Parameters".

The scope of the parameters spans implementation and operations, year by year over the specified lifecycle of the data service provider, and includes cost elements as well as workload factors and high level system configuration information.

The implementation and operations parameters will be broken down into outputs to be provided by the model, internal (derived) parameters used by the model, and inputs required by the model.

The cost estimation relationships to be used by the model will be derived from information describing actual date centers or other data service providers comparable to future ESE data service providers. Raw information received from the data service providers will be mapped to the standard reference model parameter set to build the model's comparables database, so that the database will contain an internally consistent set of parameters.

The comparables database will be used to derive the cost estimating relationships (CERs) that allow estimation of the outputs given the inputs for independent cases. This will include testing the model against independent data for an actual data service provider (for whom the actual outputs are known) and eventual use of the model to estimate the costs for a putative new ESE data service provider.

Reference Model Requirements / Levels of Service

The requirements / levels of service of the reference model are presented in Working Paper 5, "General Data Service Provider Reference Model - Requirements / Levels of Service".

The general data service provider reference model will map to a general requirements template, a statement of requirements / levels of service for a generic data service provider, in which the requirements / levels of service are defined for all of the functional areas included in the model.

The requirements / levels of service are a template in that they contain placeholders for quantitative parameters that will be defined for a specific instance of a data service provider. For example, a requirement in the template might be that "the data service provider shall provide an archive capacity of [number TB]”. A data service provider of a type that would include providing an archive would have that item in its template. If the mission of the data service provider required that it archive certain data streams and generated products that would accumulate to a total volume of 100 TB, then that value would be inserted into the template, with the result being a specific requirement for that data service provider (i.e., “the data service provider shall provide an archive capacity of 100 TB”) that could then be used in the process of generating a cost estimate for the data service provider.

Reference Model Subsets - Logical Data Service Provider Types

The general data service provider reference model includes all functions / areas of cost that a generic data service provider might perform. While an actual working data service provider could conceivably perform all of the functions included in the model, most if not all actual data service providers perform a subset of them, e.g. most providers will not have a requirement in the area of instrument / mission operations. Many well known actual data centers such as the NASA Distributed Active Archive Centers (DAACs) or the NOAA national data centers perform a subset of the general set of functions. Some data service providers, e.g. MODAPS (the MODIS Adaptive Processing System) as a sample of a science team processing facility that does not perform archive or general user distribution), are different in function from many well known data centers but fit within the framework of the data service provider reference model.

The Cost Estimation Tool will allow a planner (for example) planning a data service to support a flight project, to:

1. select those functions that are required for his/her particular mission (in effect to create a ‘custom’ subset of the general model);

2. specify the particular mission requirements the real instantiation of it must meet (e.g. data volumes to be ingested, processed, stored, and/or distributed);

3. produce an estimated cost for implementing and operating it.

A set of ‘logical data service provider types’ has been defined to enable overall ESE data service architecture studies (where a ‘data service architecture’ is a collection of data service providers and the interconnections between them), and as an option available for use by planners of individual data service provider activities. Each of type is a functional subset of the general reference model organized around a defined class of ESE role or mission. These are ‘logical’ types in that there is no explicit or implicit 1:1 mapping of an instance of a logical data service provider type to a physical entity. While some actual data service providers might match a logical type, most will perform the functions of more than one logical type, and may also perform multiple data service activities within the scope of a type (such as a DAAC that performs archive and distribution for several flight projects). Because the logical data service provider types are only a few of the possible subsets of the general model, they constitute an open set to which additions (and subtractions) can be readily made as needed to facilitate architecture trade studies or other uses.

The current set of logical data service provider types is described in Working Paper 6, "ESE Logical Data Service Provider Types", which describes each type and indicates the subset of the functional areas and requirements / levels of service that apply to it.

General DSP Reference Model - Functional Areas

This section describes the functional areas / areas of cost that comprise the Data Service Provider Reference Model. They describe the full range of functions of an abstract general data service provider. It is unlikely that an actual ESE data service provider would perform in all of the functional areas; different ones would perform in different subsets of the full set, and would perform at different levels (i.e. provide different levels of service) within functional areas.

The functional areas are primarily focused on operating activities of the data service provider. The data service provider also has additional responsibilities that require high levels of expertise in science in the discipline(s) supported by the provider, data management expertise, and information technology expertise. The Management area (see Section 3.13 below) includes lead, site-level responsibilities in these areas, and the Technical Coordination area (see Section 3.11 below) includes coordination with other ESE data service providers and broader communities in these areas.

The intent is not to provide exhaustive descriptions in great detail of every possible aspect of each of the functional areas, but rather to describe key aspects of each that are of greatest concern to either users or data service provider operators or planners or significant cost drivers.

The following sections present working definitions of the functional areas that make up the data service provider reference model.

Ingest

The ingest functional area includes receiving, reading, quality checking, cataloging, of incoming data (including metadata, documentation, etc.) to the point of insertion into the archive. Ingest can be manual or electronic with manual steps involved in quality checking, etc.

Incoming data can be received from external sources or internally generated. Ingest can include format conversion, metadata extraction, or other preparation of incoming data for archive or use within the data service provider. Ingest includes verifying that all data made available for ingest has been successfully ingested, with exceptions tracked and accounted for. Ingest must be accomplished in a timely manner as needed to meet mission requirements of the data services provider.

Processing

The processing functional area includes the generation and quality checking of new derived data products from data or products that have been ingested, or previously generated, generally on a routine, operational basis. Operational processing can be on demand as well as scheduled. Operationally generated products are often ‘standard products’ characterized by a peer reviewed, validated, reasonably stable, ‘science quality’ processing algorithm.

Processing includes ad hoc, non-operational generation of products that can include responding to requests for data mining or generation of special subsets. Processing includes process control (production planning, scheduling, monitoring, etc.) as well as product generation per se. Processing also includes reprocessing of new versions of previously generated products, either according to a reprocessing schedule or plan, or as allowed within a specified overall reprocessing capacity.

Where science or applications needs require simultaneous measurements from multiple instruments, processing performed by a data service provider can include data integration - mapping parameters from different sources to a common spatial / temporal base.

Processing can also include ‘data mining’, where software may search through many of the holdings of a data service provider for items meeting certain criteria.

The data service provider may receive the software that embodies product generation algorithms from outside developers (e.g. some Terra instrument teams for the DAACs currently) who are responsible for the initial delivery and for delivering updated versions. Where quality, especially science quality, of products remains the responsibility of an outside developer, processing includes supporting quality checking by the science software developer. Support provided by the data service provider for integration and test of this ‘science software’ is included as an activity under processing. In cases where a data service provider develops algorithm software, that effort (i.e. development, integration, and test) is included under Implementation.

The data service provider may also accept software from science or applications users to produce a research product, perform data integration, or perform data mining.

Documentation

The documentation functional area includes the development (or upgrading of received) data and product documentation (including user guides, catalog interfaces, etc.) to meet SEEDS adopted documentation standards, including catalog information (metadata), user guides, etc., through consultation with data providers, algorithm developers, flight projects, etc. Knowledge capture is a critical concern - the data service provider must be committed to pro-actively capture knowledge of instruments, calibration, processing history, etc., from its data sources (e.g. instrument teams).

SEEDS adopted documentation standards may include FGDC (Federal Geographic Data Committee) metadata standards, documentation standards for long term archiving, Algorithm Theoretical Basis Documents (or equivalent, which must reflect ‘as-built’ algorithms), Data Software Interface Specifications, etc. When science needs require that multiple versions of a product be held, the documentation of each version must include the provenance information (e.g. processing algorithm) peculiar to it.

Documentation should include comments received from users on their experience with the data and products (product accuracy, usability, etc.), perhaps in the form of FAQ's (Frequently Asked Questions) for products, both from scientists on staff or working closely with the provider and from the general user community.

Documentation should include read software and other appropriate tools for data access kept current with commonly available technology. Documentation includes maintenance and refresh according to best industry practice or SEEDS policy.

Documentation needs will evolve, e.g. information relevant to intellectual property rights may be needed.

Archive

The archive functional area includes the insertion of data into archive storage, and data stewardship - management, handling and preservation of data, metadata, and documentation within a data service provider's archive. Inserted data can include data ingested from sources external to the site, or data/products generated on-site.

Data stewardship / preservation includes quality screening of data entering and exiting the archive, quality screening of archive media, tested and verified backup and restoration capability, and accomplishing migrations from one type of media to another.

Insertion into the archive can be electronic or manual (e.g. hanging tapes on a rack or popping them into a robotic silo).

Search and Order

The search and order functional area includes providing access to catalog information (a range of descriptive information to aid in selecting data and products) and a search and order capability to users, and receiving user requests for data.

"Search and order" in this context is used in a very broad sense; search and order includes support for system to system interactions as well as conventional search and order by users directly. For example, system to system interactions might include a program running on a user platform accessing the data service provider system directly, locating a needed product, and executing a protocol (e.g. for user registration, security) to gain access to it.

"Search", whether by a user directly or through a system-system interaction, implies applying criteria that might include geophysical parameter(s), spatial-temporal coverage, specific product names, etc., to the metadata describing available data and products and returning to the user listings supplemented by descriptive information of those data or product types and instances that meet the criteria.

“Order” implies a request/permission step, regardless of how implemented (e.g. manual or automated), where a request for a set of data or product instances, perhaps the results of (or a selected subset of the results of) a search, is processed and accepted or denied.

Search and order can include providing local user interface and capability and/or providing an interface to a broader based, cross-site search and order capability (e.g. DAACs supporting search and order via the EOS Data Gateway).

Access and Distribution

The access and distribution functional area includes fetching the requested data from the archive, performing any subsetting, resampling, reformatting / format conversion (e.g. to a GIS (Geographic Information System) format), reprojection, or packaging, and providing the end product to the user by electronic means or on physical media.

“Access” is included to embrace a service allowing a program running on a user platform to access data and products from the data service provider directly, through an appropriate protocol, perhaps as a seamless extension of the system to system search and order described above.

Access and distribution can be performed on an operational basis, meaning in part that a data service provider will formally commit to terms of service in a level of service agreement or equivalent.

Access and distribution is an area likely to see substantial evolution in the next five to ten years, perhaps especially if distributed computing comes into play on a significant scale. Highly automated access techniques, software agents, and new tools for data discovery, access, integration from multiple sources, etc., will become available.

Note: Success from a user point of view may be even more dependent on a product’s format than the speed of its delivery (what if a product is delivered in 30 seconds but the form is such that a user needs to spend several hours to be able to use it, vs a product delivered in 30 minutes but in a form that can be used directly?). The data service provider should take care to offer formats (whether as a default or an option) that are directly useable by the largest possible fraction of its user community.

User Support

The user support functional area includes support provided in direct contact with users by user support staff, including responding to queries, taking of orders, staffing a help desk (i.e., staff awaiting user contacts who can assist in ordering, track and status pending requests, resolve problems, etc.), etc. User support staff includes science expertise to assist users in selecting and using data and products.

The demands on user support will increase with the proliferation of data types, data sources, and tools for users, continuing or increasing the need for highly trained user support staff even as user interactions become more automated and more automated user support aids become available (beginning with on-line documentation, FAQ, etc.).

User support also includes outreach to potential new users and education / training for current or potential new users.

User support should also be a channel for feedback from the users to the data service provider, whether comments on particular data or products or on the provider’s services and support.

User support includes coordination of user support guidelines and practices across the network of ESE data service providers and with other data centers as needed to support the ESE science and applications program - see Technical Coordination.

Instrument / Mission Operations

The instrument / mission operations functional area includes monitoring instrument and spacecraft performance, generating instrument and spacecraft commands, and event scheduling (using NASA or other appropriate operational mission management services).

Sustaining Engineering

Sustaining engineering includes maintenance and enhancement of custom applications software (including any science software embodying processing algorithms developed by the site).

Engineering Support

Engineering support includes some or all of the following as applicable at a particular site: systems engineering, test engineering, configuration management, coordination of hardware maintenance by vendors, COTS procurement, installation of COTS upgrades, system administration, database administration, network/communications engineering, and security.

Engineering support is internal, directed toward the internal operation of the data service provider.

Technical Coordination

Technical coordination includes participation in SEEDS system level processes, including coordination on data management, data stewardship (including standards for content of life cycle data management plans), standards and best practices (including quality assurance standards and practices), interfaces, common metrics, and interoperability (e.g. for data access and integration), across / within SEEDS and with other systems and networks as needed to support the ESE program.

This area includes coordination on evolution of the overall ESE data service architecture (including an examination of the changing needs of the ESE science and applications program and the consequent impacts on the roles, missions, and services of ESE data service providers).

Technical coordination includes participation in SEEDS system level processes to coordinate user support guidelines and practices across the network of ESE data service providers and with other data centers as needed to support the ESE science and applications program.

Technical coordination includes participation in SEEDS level and/or bilateral processes to coordinate production and delivery of products between ESE data service providers.

Technical coordination includes cooperating with other ESE data service providers in representing ESE / SEEDS in broader community processes in areas such as standards, interoperability, data management, security, etc.

Technical coordination, which by its nature includes engineering, is directed outward, supporting the data service provider as one element of a system of cooperating centers.

Implementation

Implementation includes development of, and making operational, the data and information system capabilities required by the data service provider to perform its mission, including design and implementation of the data system (hardware and system software) and applications software. Implementation can recur during the operating period as systems are expanded or replaced.

In addition to a major implementation effort, implementation can include ongoing applications software development. Implementation can include development of software tools for use by users to unpack, subset, or otherwise manipulate products provided by the data service provider.

In some cases applications software will include product generation software embodying science algorithms, e.g. to produce a product to meet a particular user need. Applications software can include software to perform a ‘data mining’ or data integration operation to meet a user need.

Management

Management includes management and administration at the data service provider level (“front office”) and direct management of functional areas. Management also includes staff with overall responsibility for internal and external science activities, information technology planning, and data stewardship.

Management includes planning information technology upgrades / technology refreshes, based on assessments of changing mission or user needs and availability of new technology.

Management includes developing data stewardship practices, performing data administration with science advice (via the User Advisory Group and other appropriate bodies), developing and maintaining life cycle data management plans (which address data migrations).

Management also includes coordinating the science activities within the data service provider and its interaction with the ESE and broader science community, including a visiting scientist program, collaboration among ESE data service providers to support science needs, annual Enterprise peer review, and support for its User Advisory Group (which includes representation from the science, applications, education, etc., communities as appropriate for a given data services provider) and any other ESE or broader advisory activities that may be appropriate.

Management also includes participation in SEEDS management processes, strategic planning, coordination with other data centers and activities beyond ESE/SEEDS.

Management includes performing supervisory, financial administration, and other administrative functions.

Facility / Infrastructure

Facility / Infrastructure includes provision and maintenance of a fully furnished and equipped, environmentally controlled, physically secure facility to house data service provider staff, systems, and data and information holdings, including a backup facility for its data and information holdings. An off-site backup facility would be one sufficiently removed from the data service provider's primary site such that a fire, tornado, or other event that destroys the primary site would be very to extremely unlikely to also destroy the backup site (a risk analysis would be performed on a site by site basis).

This area includes resource planning, logistics, supplies inventory and acquisition, and facility management.

This area includes maintenance of system and site security according to established NASA security policies and practices.

Facility/Infrastructure also includes a variety of non-staff cost factors such as supplies, facility lease and utility costs and similar overhead costs, hardware maintenance, COTS licenses, etc.

References and Acronyms

The References Section and the Acronym List for all of these Working Papers is in the document

"References and Acronyms for the Levels of Service / Cost Estimation Working Papers ".

SEEDS

Working Paper Four: Data Service Provider Model, Model Parameters

April 24, 2002

G. Hunolt, SGT, Inc.

Outline

Introduction

This working paper is the fourth of a set of papers that describes the SEEDS (Strategic Evolution of Earth Science Enterprise Data Systems) Levels of Service (LOS) / Cost Estimation (LOS/CE) study. The study goal is to develop a cost estimation model and coupled requirements and levels of services to support the SEEDS Formulation team in estimating the life cycle costs of future ESE data service providers and supporting systems, where 'data service provider' is used as a generic term for any data/information related activity. The set of working papers is intended to serve as a vehicle for coordinating work on the project, obtaining feedback and guidance from ESDIS SOO and the user community, and as embryos of reports that will be produced as the task proceeds.

As working papers, each version of each working paper that appears represents a snapshot in time, with the work in various stages of completion. As work progresses the content (and sometimes the organization) of the working papers will change reflecting progress made, responses to feedback and guidance received, etc.

This fourth working paper of the set will define and describe the parameters of the general data service provider reference model developed for the LOS/CE study and their relationship with the model's requirements / levels of service. The paper reflects results of the February, 2002, SEEDS Community Workshop. The parameter list and definitions can be expected to undergo considerable evolution as work on developing the model and building the comparables database proceeds over the life of the project.

Section 2 describes the Data Service Provider Reference Model and shows how the requirements / levels of service and model parameters are integral components of the model, organized around the model's functional areas. Section 3 presents the model parameters organized by functional area. Section 4 presents a mapping of the model parameters and the requirements / levels of service.

Data Service Provider Reference Model Parameters

This section describes the Data Service Provider Reference Model, a functional model of a generic data service provider.

The reference model has three integrated components:

1) A set of 'functional areas' (see Working Paper 3, "Data Service Provider Reference Model - Functional Areas") that collectively comprise the full range of functions that a generic data service provider might perform and the areas of cost that must be considered by the cost estimation by analogy model.

2) A set of requirements and levels of service for each functional area (see Working Paper 5, “Data Service Provider Reference Model - Requirements / Levels of Service”).

3) A set of parameters (defined below) for each functional area that constitute a quantitative description of the workload, staff effort, and any other factors that contribute to cost for that area, additional ‘roll-up’ parameters that sum items such as staff effort across the functional areas, and other parameters like labor rates that are required for cost estimation.

These three aspects of the model are closely coupled to ensure the internal consistency of the model. The set of functional areas is the underpinning; both the model parameters and requirements / levels of service are organized according to the functional areas. The requirements / levels of service and the model parameters are coupled in that the definitions of the requirements / levels of service embody model parameters. This integration of the three components of the model is intended to ensure that estimated costs are driven by and traceable to requirements to the fullest extent possible.

The scope of the reference model parameters spans implementation and operations, year by year over the specified lifecycle of the data service provider, and include cost elements as well as workload factors and high level system configuration information.

The implementation and operations parameters will be broken down into outputs to be provided by the model, internal (derived) parameters used by the model, and inputs required by the model.

The cost estimation relationships to be used by the model will be derived from information describing actual date centers or other data service providers comparable to future ESE data service providers. Raw information received from the data service providers will be mapped to the standard reference model parameter set to build the model’s comparables database, so that the database will contain an internally consistent set of parameters.

The comparables database will be used to derive the cost estimation relationships that allow estimation of the outputs given the inputs for independent cases (i.e. testing against independent data for an actual data service provider and eventual use of the model to estimate the costs for a putative new ESE data service provider).

Reference Model Parameters and Requirements / Levels of Service

As noted, the general data service provider reference model includes a general requirements template, a statement of requirements / levels of service for a generic data service provider, in which the requirements / levels of service are defined for all of the functional areas included in the model.

The requirements / levels of service are a template in that they contain placeholders for quantitative parameters that will be defined for a specific instance of a data service provider. For example, a requirement in the template might be that “the data service provider shall provide an archive capacity of [number TB]". A data service provider of a type that would include providing an archive would have that item in its template. If the mission of the data service provider required that it archive certain data streams and generated products that would accumulate to a total volume of 100 TB, then that value would be inserted into the template, with the result being a specific requirement for that data service provider (i.e., "the data service provider shall provide an archive capacity of 100 TB") that could then be used in the process of generating a cost estimate for the data service provider.

The requirements / levels of service template contains reference model parameters, or place-holders for parameters that must be supplied by the user of the cost estimation tool that is built on the reference model. For example, quantities that are defined in the levels of service associated with a requirement are model parameters whose values are given - the user selects the one applicable for his or her specific case. As a second example, the ingest requirement contains placeholders for the numbers of product types, instances of each type, volume, etc., to be ingested. These are all model parameters, which the user provides as input when using the cost estimation tool. Other parameters are not contained either directly or as place holders in the requirements / levels of service. These include control parameters such as an annual inflation rate to be assumed, which must be specified by the cost estimation tool user, or parameter that are the cost estimation tool's output, such as ingest operator FTE, which would be computed by the model based on ingest workload parameters provided by the user and a cost estimating relationship, or internal parameters that are intermediate steps between the inputs and the outputs.

Section 2 below defines and describes the reference model parameters.

Section 3 presents a mapping of the Data reference model parameters defined in Section 2 to the requirements / levels of service. The intent is to show which parameters fall within the scope of each requirement, and to ensure that each requirement / levels of service that should have one or more parameters associated with it actually does. Second and third level derived parameters (i.e. parameters internal to the model) are not shown.

Data Service Provider Reference Model Parameter Definitions

This section presents the definitions of the parameters used by the reference model.

Introduction

This section introduces the description and definition of reference model parameters that follows in sections 2.2, 2.3, and 2.4.

The reference model parameters are a standard set of parameters that includes some that cover a data service provider as a whole and some that are mapped to the model's functional areas as they apply (i.e., not all parameters are applicable to all functional areas).

The scope of the parameters spans implementation and operations, year by year over the specified lifecycle of the data service provider, and include cost elements as well as workload factors and high level system configuration information.

The implementation and operations parameters will be broken down into outputs to be provided by the model, internal (derived) parameters used by the model, and inputs required by the model.

The cost estimation relationships to be used by the model will be derived from information describing actual date centers comparable to future ESE data service providers. As was done for the Best Practices / Benchmark Study, raw information received from the data service providers will be mapped to the standard reference model parameter set to build the model's database, so that the model's database will contain the same set of output, input, and derived internal parameters covering implementation and operation as will be used for cost estimation. This is necessary, since the model database will be used to derive the cost estimation relationships that allow estimation of the outputs given the inputs for independent cases (i.e. testing against independent data for an actual data service provider and use of the model to estimate the costs for a putative new ESE data service provider).

Implementation includes capital and staff costs associated with developing, implementing, integrating and testing the data service provider's data and information system, and facility start-up / preparation costs. Implementation is assumed to be spread over a specified number of years. Implementation can overlap the start of operations. Implementation can also recur during the operating period, e.g. allowing for 'technology refresh'.

Operation includes hardware maintenance, sustaining engineering, operations staff, supplies (e.g. storage and archive media), recurring facility costs, etc.

The parameters are defined in Section 2.2 grouped by the reference model's functional areas (see White Paper 3, "Data Service Provider Reference Model - Functional Areas". Within each functional area group, the parameters are sorted into internal derived parameters used by the model, input parameters that must be provided by a user of the cost estimation tool, and output parameters, i.e. required outputs from the cost estimation tool.

Section 2.3 contains a list of the cost estimation model output parameters, and Section 2.4 contains a list of the user input parameters required to run the cost estimation model. Both lists are drawn from the parameters defined in Section 2.2.

Reference Model Parameter Definitions

This section contains a master list of the data service provider reference model parameters and their definitions. The list is grouped by the functional areas described in Section 4.1, followed by facility / infrastructure parameters and data service provider level parameters (some of which are roll-ups from the preceding functional areas). Within each functional area, the parameters will be sorted between internally computed parameters, parameters provided as user input when executing the cost estimation model, and cost estimation model output parameters.

Information included about each parameter is:

- Parameter Name;
- Parameter Definition;
- Reference to Requirements / Levels of Service (provided in Working Paper 5, "Data Service Provider Reference Model - Requirements / Levels of Service"). The reference will be the number, in brackets, of the sub-section within Working Paper 5 that contains the requirement.

Ingest

These parameters describe or relate to the ingest of data and products into the data service provider from external sources / providers.

2.2.1.1 Internal Computed Parameters

1. **Total Ingest FTE.** The total estimated annual FTE (Full Time Equivalent) effort for the Ingest functional area, including any effort in addition to actual operational effort.
2. **Ingest Management FTE.** Includes direct management associated with the Ingest functional area. Computed from technical and operations staffing.
3. **Ingest Technical FTE.** Includes ingest technical staff exclusive of direct operations staff.
4. **Ingest Ops FTE**. The estimated annual FTE effort for direct operational activity (e.g. computer operators, ingest technicians).

5. **Ingest Volume/Yr.** The annual volume of data and/or products that are ingested by the site. {2.1 a}
6. **Ingest Volume/Yr per FTE.** The annual volume divided by the total staff effort for the Ingest functional area.
7. **Ingest Volume/Yr per Ops FTE.** The annual volume divided by the direct operations staff effort for the Ingest functional area.
8. **Product Types Ingested/Yr.** The annual number of different product types ingested (i.e. data streams ingested) from external sources by the site. {2.1 a}
9. **Product Ingest Formats/Yr.** The number of distinct different product or data formats handled by the Ingest functional area. {2.1 a}
10. **Products Ingested/Yr.** The annual number of products ingested from external sources by the site. { 2.1 a}
11. **Products Ingested/Yr per FTE.** The annual products ingested count divided by the total staff effort for the Ingest functional area.
12. **Products Ingested/Yr per Ops FTE**. The annual products ingested count divided by the direct operations staff effort for the Ingest functional area.
13. **Ingest Function LOS.** The overall measure of ingest function level of service (LOS) integrated over product types. Same values as Ingest LOS for Product Type. {2.1 a}

2.2.1.2 User Input Parameters

(a) **Ingest Product Type Name.** The name of product or data type. {2.1 a}

(b) **External Ingest Interfaces.** The number of distinct external interfaces via which data streams or products are ingested each year.

(c) **Ingest Source**. The source or provider of the product or data type. {2.1 a}

(d) **Ingest Delivery Means**. The means of delivery from the source to the data service provider (values: 1 - electronic, 2 - media).

(e) **Ingest LOS for Product Type.** Levels of service, assigned by product type, associated with the ingest function are: 1) operational (time-critical) ingest with immediate verification of data integrity and quality; 2) routine ingest and verification of data quality and integrity without tight time constraints; 3) ad hoc or intermittent ingest on a non-operational basis with verification of data quality and integrity; 4) ad hoc or intermittent ingest on a non-operational basis. {2.1 a}

(f) **Products of Type Ingested Per Day.** The typical number of instances (individual products of the type) ingested per day. {2.1 a}

(g) **Volume of Type Ingested Per Day.** The average data volume ingested per day for this data or product type. {2.1 a}

(h) **Ingest Product Type Format.** Incoming format for product type. {2.1 a}

(i) **Conversion Format for Product Type.** The format into which instances of the product type are converted to on ingest, if applicable. {2.1 a}

(j) **Ingest Product Type Retention Period.** The data service provider's planned retention for this data or product type, can be N years after receipt, or indefinite, for use in computing Archive Volume and Archive Products. (Should be included in applicable life cycle data management plan.) {2.1 a, 2.4 a, 2.4 b}

Processing

These parameters describe or relate to the generation of products by the data service provider.

2.2.2.1 Internal Computed Parameters

(a) **Total Processing FTE.** The total estimated annual FTE effort for the Processing functional area, including any effort in addition to actual operational effort.

(b) **Processing Management FTE.** Includes direct management associated with the Processing functional area. Computed from technical and operations staffing.

(c) **Processing Technical FTE.** Includes technical and science staff exclusive of direct operations staff. Includes staff supporting science software integration and test, cross-calibration specialists as applicable. {2.2 e, 2.2 f}

(d) **Processing Ops FTE**. The estimated annual FTE effort for direct operational activity (e.g. computer operators, production monitors).

(e) **Volume/Yr of New Operational Products.** The annual volume of operational products generated by the site. {2.2 a}

(f) **Volume/Yr of New Ad Hoc Non-Operational Products.** The annual volume of ad hoc, non-operational products generated by the site. {2.2 b}

(g) **Volume/Yr of New Products Generated.** The total annual volume of new products generated by the site. {2.2 a, 2.2 b}

(h) **Volume/Yr of Reprocessed Products Generated.** The annual volume of reprocessed products generated by the site. {2.2 c, 2.2 d}

(i) **Processing Volume/Yr.** The annual total volume of new and reprocessed data and/or products that are generated by the site. {2.2 a, 2.2 b, 2.2 c, 2.2 d}

(j) **Processing Volume/Yr per FTE.** The annual processing volume divided by the total staff effort for the Processing functional area.

(k) **Processing Volume/Yr per Ops FTE.** The annual processing volume divided by the direct operations staff effort for the Processing functional area.

(l) **New Operational Products Generated/Yr.** The annual number of new operational products generated per year by the site. {2.2 a}

(m)**New Ad Hoc Non-Operational Products Generated /Yr.** The annual number of new ad hoc non-operational products generated per year by the site. {2.2 b}

(n) **New Products Generated/Yr.** The total annual number of new products generated per year by the site. {2.2 a, 2.2 b}

(o) **Reprocessed Products Generated/Yr.** The annual number of reprocessed products generated per year by the site. {2.2 c, 2.2 d}

(p) **Product Types Generated/Yr.** The annual number of different product types generated by the site. { 2.2 a, 2.2 b, 2.2 c, 2.2 d}

(q) **Product Generation Formats/Yr.** The number of distinct different product or data formats handled by the Processing functional area.

(r) **Products Generated/Yr.** The annual total number of new and reprocessed products generated by the site. {2.2 a, 2.2 b, 2.2 c, 2.2 d}

(s) **Products Generated/Yr per FTE.** The annual products generated count divided by the total staff effort for the Processing functional area.

(t) **Products Generated/Yr per Ops FTE**. The annual products generated count divided by the direct operations staff effort for the Processing functional area.

(u) **Operational Processing LOS.** The overall measure of operational processing level of service integrated over product types. Same values as Operational Processing LOS for Type. {2.2 a}

(v) **Non-Operational Processing LOS.** The overall measure of ad hoc, non-operational processing level of service integrated over product types. Same values as Operational Processing LOS for Type. {2.2 b}

(w) **Reprocessing Aggregate Capacity LOS.** The measure of overall capacity for reprocessing. Same values as Reprocessing Capacity for Type. {2.2 c}

2.2.2.2 User Input Parameters

(a) **Product Type Name.** The name of product type. {2.2 a, 2.2 b, 2.2 c, 2.2 d}

(b) **Product Type Software Source.** A flag, for each product type, that indicates whether the algorithm software produced in-house or received from another activity. {2.2 e}

(c) **Product Type QA Function.** A flag that indicates whether the quality assurance (QA) is an in-house function or whether another activity involved.

(d) **Production Mode for Type.** Indicates whether this product type is produced operationally or on an ad hoc, non-operational basis. {2.2 a, 2.2 b, 2.2 c, 2.2 d}

(e) **Operational Production Mode for Type**. Is the operational generation of this product type performed on demand, or on a routine, scheduled basis. {2.2 a}

(f) **Products of Type Generated per Day.** The typical number of instances (individual products of the type) generated per day. {2.2 a, 2.2 b, 2.2 c, 2.2 d}

(g) **Volume of Type Generated per Day.** The average data volume generated per day for this product type. {2.2 a, 2.2 b, 2.2 c, 2.2 d}

(h) **Product Type Format.** The format in which the new products are produced. {2.2 a, 2.2 b}

(i) **Generated Product Type Retention Period.** The data service provider’s planned period of retention for this product type (i.e. for each new version that is generated), can be N years after production, or indefinite, or by a rule (e.g. delete if reprocessed, or keep N versions). {2.2 a, 2.2 b, 2.4 a}

(j) **Reprocessing Capacity for Type.** The data service provider's required reprocessing capacity for this product, as a multiple of the original processing rate. This is the level of service reprocessing of standard products, values: 1 - the capacity for reprocessing shall be 9 times the original processing rate; 2 - 6 times; 3 - 3 times. {2.2 c}

(k) **Reprocessing Plan for Type.** The nominal interval in years at which the data service provider would reprocess the instances of the product type (i.e. create new versions of product instances). {2.2. d}

(l) **Operational Processing LOS by Type.** Level of service associated with operational processing of a given product type, values: 1 - standard products shall be generated within 2 days of ingest/availability of required inputs, 2 - within 7 days, 3 - within 30 days. {2.2 a}

(m)**Non-Operational Processing LOS by Type.** Level of service associated with ad hoc, non-operational processing of a given product type, values: 1 - specific targets for processing adopted on a case by case basis; 2 - general goals for processing; 3 - no goals, purely ad hoc processing. {2.2 b}

(n) **Reprocessing LOS by Type.** Level of service associated with reprocessing according to a schedule (see Reprocessing Plan for Type), values: 1 - reprocess according to negotiated schedule; 2 - reprocess to meet general goals of schedule; 3 - reprocess on time available basis to intent of schedule. {2.2 d}

(o) **Science Software LOS.** Level of Service associated with acceptance of science algorithm software from users, values: 1 - accept standard (operational), research product generation software, and/or data integration and data mining software; 2 - accept research product generation software, and/or data integration and data mining software, 3 - accept standard (operational) or research product generation software; 4 - accept research product generation software; 5 - accept standard (operational) product generation software. {2.2 e}

(p) **Cross-Calibration Flag.** Indicates if data service provider requires technical expertise in producing products from multiple inputs (e.g. a time series from data collected by a series of instruments on successive platforms) requiring cross-calibration, etc. {2.2 f}

Documentation

These parameters describe or relate to the generation, or bringing up to standard, by the data service provider of documentation of data and products, where 'documentation' includes all descriptive information such as catalog metadata as well as user guides, format descriptions, etc.

2.2.3.1 Internal Computed Parameters

(a) **Total Documentation FTE.** The total estimated annual FTE effort for the functional area, including any effort in addition to actual operational effort.

(b) **Documentation Management FTE.** Includes direct management associated with each functional area. Computed from technical staffing.

(c) **Technical FTE.** Includes technical staff working on documentation (including metadata) review, creation, and update.

2.2.3.2 User Input Parameters

(a) **Documentation LOS.** Documentation level of service, values: 1- data and product holdings documented to the standard for long term archiving; 2 - documentation ensured to be sufficient for current use; 3 - documentation only as received from product provider. {2.3 a}

(b) **User Comment LOS.** Level of service for incorporating user feedback on products into product documentation. Values: 1) data and products routinely updated with user comments; 2 - data and products occasionally updated with user comments; 3 - data and products rarely updated with user products. {2.3 b}

(c) **DIF's Delivered/Yr.** A count of the number of Directory Interchange Format (DIF) records provided by the site to the Global Change Master Directory. {2.3 c}

Archive

These parameters describe or relate to the archiving of data and products by the data service provider.

2.2.4.1 Internal Computed Parameters

(a) **Total Archive FTE.** The total estimated annual FTE effort for the Archive functional area, including any effort in addition to actual operational effort.

(b) **Archive Management FTE.** Includes direct management associated with each functional area. Computed from technical and operations staffing.

(c) **Archive Technical FTE.** Includes technical staff exclusive of direct operations staff.

(d) **Archive Ops FTE**. The estimated annual FTE effort for direct operational activity (e.g. computer operators).

(e) **Archive Insert Volume/Yr.** The annual volume of data and/or products that are inserted into the site's archive. {2.4 a, 2.4 b}

(f) **Archive Insert Volume/Yr per FTE.** The annual Archive Insert Volume divided by the total staff effort for the Archive functional area.

(g) **Archive Insert Volume/Yr per Ops FTE.** The annual Archive Insert Volume divided by the direct operations staff effort for the Archive functional area.

(h) **Product Types Archived/Yr.** The annual number of different product types added to the site's archive. {2.4 a, 2.4 b}

(i) **Product Archive Formats/Yr.** The number of distinct different product or data formats handled by the Archive functional area.

(j) **Products Archived/Yr.** The annual number of products added to the site's archive. {2.4 a, 2.4 b}

(k) **Products Archived/Yr per FTE.** The annual products archived count divided by the total staff effort for the Archive functional area.

(l) **Products Archived/Yr per Ops FTE**. The annual products archived count divided by the direct operations staff effort for the Archive functional area.

(m)**Primary Archive Volume.** The year by year cumulative total volume of data contained in the site's primary archive. {2.4 a, 2.4 b}

(n) **Backup Archive Volume**. The year by year cumulative volume of data contained in the site's backup archive. {2.4 h}

(o) **Archive Volume.** The year by year total cumulative volume of data contained in the site's primary and backup archives. The sum of Primary Archive Volume and Backup Archive Volume. {2.4 a, 2.4 b, 2.4 h}

(p) **Archive Volume per FTE.** The archive volume divided by the total effort for the archive functional area.

(q) **Archive Volume per Ops FTE.** The archive volume divided by the direct operations staff effort for the Archive functional area.

(r) **Archive Media Units**. The number of media units (e.g. tapes) required to hold the data contained in the site's archive.

2.2.4.2 User Input Parameters

(a) **Archive Media Type**. The archive media type(s) used by the data service provider. [Background];

(b) **Archive Media Standard.** The standard that this media type is in compliance with, or none, level of service values: 1 - archive media consistent with best commercial practice; 2 - archive media and system vendor independent; 3 - archive media vendor independent. {2.4 i}

(c) **Archive Media Unit Capacity.** The volume of data that can be written to a single unit of the archive media type.

(d) **Archive Media Fill Rate.** The average or typical fraction of a single archive media unit of the archive media type that is filled with archived data or products.

Note: Have to allow for multiple archive media types. Items 1, 3 and 4 above are used in conjunction with Archive Volume to project Archive Media Units.

(e) **Archive Capacity.** The maximum capacity of the site's primary archive storage, as either indefinite (i.e. a function of the retention plans without an arbitrary limit, or limited by a specified upper bound. This is the archive capacity level of service, values: 1 - archive capacity is cumulative sum of all data ingested plus all products generated, less deletions per retention plans; N - archive capacity is limited to a specified threshold of N (year by year values). {2.4 e}

(f) **Archive Backup LOS.** The level of service associated with archive backup by the site, values: 1- full off-site backup, with regular sampling to verify integrity; 2 - partial, [Backup Fraction - % of archive backed up], off-site backup, with sampling; 3 - partial, [Backup Fraction - % of archive backed up], on-site backup with sampling. {2.4 h}

(g) **Archive Backup Fraction.** The fraction of the Primary Archive Volume that is to be backed up. {2.4 h}

(h) **Archive Backup Plan**. The data service provider's plan for backing up its archive, including the fraction of the primary archive that is backed up - copied to storage media, and whether the backup storage is on-site or off-site. Level of service, values: 1 - full off-site backup, with sampling to verify integrity; 2 - partial backup, off-site, with sampling; 3 - partial backup, on-site, with sampling. {2.4 c, 2.4 h }

(i) **Archive Migration Plan.** The plan that the data service provider has to migrate its archive to a new media and/or archive system, including the period in years between migrations and the migration rate. Includes level of service, values: 1 - planned migration; 2 - no planned migration, but ad hoc migration as need is seen to arise. {2.4 j}

(j) **Archive Monitoring.** Archive quality monitoring to support preservation; the fraction of the archive that is scanned for media integrity per year. Level of service values: 1 - 10% per year random screening; 2 - 5% per year random screening; 3 - 1% per year random screening. {2.4 c, 2.4 g}

(k) **Archive Entry/Exit Screening.** Archive entry and/or exit data quality screening, level of service values: 1- exit and entry screening; 2 - entry screening. {2.4 c, 2.4 f}

Search and Order

These parameters describe or relate to catalog search and order, allowing users to search metadata for, identify, and request products.

2.2.5.1 Internal Computed Parameters

(a) **Total Search and Order FTE.** The total estimated annual FTE effort for the Search and Order functional area, including any effort in addition to actual operational effort.

(b) **Search and Order Management FTE.** Includes direct management associated with the Search and Order functional area. Computed from technical and operations staffing.

(c) **Search and Order Technical FTE.** Includes technical staff exclusive of direct operations staff.

(d) **Search and Order Ops FTE**. The estimated annual FTE effort for direct operational activity (e.g. computer operators).

(e) **Internal Catalog Size.** Internal catalog search and order function size - number of product instances included in the catalog.

2.2.5.2 User Input Parameters

(a) **Search and Order Scope.** A level of service parameter that establishes the scope of the search and order service offered by the site. Values: 1 - public access to all users; 2 - access to the science and applications community; 3 - access to a limited team of scientists. {2.5 a}

(b) **Internal Catalog Search Complexity.** The complexity of the search capability offered to the user, a level of service parameter, values: 1 - search for instances of multiple product types that pertain to a specified object or phenomenon; 2 - search for instances of product types by geophysical parameter, time, and space across multiple product types; 3 - search for instances of multiple product types by time and space (coincident search); 4 - search for instances of single product type by time and space; 5 - search for instances of a product type from a list of instances available. {2.5 b}

(c) **External Catalog Search and Order.** The type of interface, if any the data service provider provides to an external search and order capability, values: 1 - none, 2 - external user interface client accesses local catalog information, provides user requests to data service provider, 3 - local catalog information provided to external catalog system which provides user requests to data service provider. {2.5 e}

(d) **Descriptive Information LOS.** A level of service parameter that establishes the type of descriptive information to be available for product types or instances returned by a search, values: 1 - detailed algorithm and use explanations, references to papers, standard guide and DIF metadata; 2 - references to papers, standard guide and DIF metadata; 3 - standard guide and DIF metadata {2.5 c}

(e) **System-System Search.** A flag that indicates the presence (1) or absence (0) of an automated system-system search capability. {2.5 d}

Access and Distribution

These parameters describe or relate to providing access to and/or distribution of products to users, either on an operational basis or in response to user requests (a.k.a. ‘ad hoc’). This includes providing automated ‘system-system’ access.

2.2.6.1 Internal Computed Parameters

(a) **Total Access and Distribution FTE.** The total estimated annual FTE effort for the Access and Distribution functional area, including any effort in addition to actual operational effort.

(b) **Access and Distribution Management FTE.** Includes direct management associated with the Access and Distribution functional area. Computed from technical and operations staffing.

(c) **Access and Distribution Technical FTE.** Includes technical staff exclusive of direct operations staff.

(d) **Access and Distribution Ops FTE**. The estimated annual FTE effort for direct operational activity (e.g. computer operators, distribution technicians).

(e) **Access and Distribution Volume/Yr.** The annual volume of data and/or products that are distributed by the site.

(f) **Distribution Volume/Yr per FTE.** The annual distribution volume divided by the total staff effort for the Distribution functional area.

(g) **Distribution Volume/Yr per Ops FTE.** The annual distribution volume divided by the direct operations staff effort for the Distribution functional area.

(h) **Product Types Distributed/Yr.** The annual number of different product types distributed by the site.

(i) **Product Distribution Formats/Yr.** The annual number of distinct different product or data formats handled by the Distribution functional area. {2.6 b}

(j) **Product Types/Yr Distributed Operationally.** The annual number of product types distributed on an operational basis - on a schedule or by rule to specified users.

(k) **Product Formats/Yr Operational.** The annual number of different product formats used for products distributed operationally.

(l) **Network Products/Yr Operational.** The annual number products distributed operationally by network.

(m)**Network Volume/Yr Operational.** The annual volume of data/products distributed operationally by network.

(n) **Media Products/Yr Operational**. The annual number products distributed operationally by media.

(o) **Media Volume/Yr Operational.** The annual volume of data/products distributed operationally by media.

(p) **Product Formats/Yr By Request.** The annual number of different product formats distributed by in response to user request.

(q) **Products Distributed/Yr.** The annual number of products distributed by the site.

(r) **Products Distributed/Yr per FTE.** The annual products distributed count divided by the total staff effort for the Distribution functional area.

(s) **Products Distributed/Yr per Ops FTE**. The annual products distributed count divided by the direct operations staff effort for the Distribution functional area.

(t) **Network Distribution Volume/Yr.** The annual volume of data distributed by the site by network, usually by FTP (File Transfer Protocol). {2.6 f}

(u) **Network Distribution Products/Yr.** The annual number of products distributed by the site by network. {2.6 f}

(v) **Media Distribution Volume/Yr.** The annual volume of data distributed by the site on media. {2.6 g}

(w)**Media Distribution Products/Yr.** The annual number of products distributed by the site on media. {2.6 g}

(x) **Distribution Media Units/Yr**. The annual number of media units (i.e. the sum of the number of tapes of various sorts, CD-ROMs, DVDs, etc., used for distribution by the site). {2.6 h}

(y) **Distribution Media Types/Yr.** The types of distribution media used by the site (CD-ROM, DVD, 8mm tape, etc.). {2.6 h}

(z) **Transmigration Products/Yr.** The number of products per year to be migrated to another center. {2.6 i}

(aa) **Transmigration Volume/Yr.** The volume of data and products to be migrated to another center. {2.6 i}

2.2.6.2 User Input Parameters

(a) **Distribution External Interfaces.** The number of distinct external interfaces for distribution, especially for operational distribution.

(b) **Access and Distribution Scope.** A level of service parameter that establishes the scope of the distribution service offered by the site. Values: 1 - public access to all users; 2 - access to the science community; 3 - access to a limited team of scientists. {2.6 a}

(c) **Access and Distribution Service Modes.** A parameter characterizing the modes of distribution service offered by the site: distribution operationally, by subscription, and/or in response to request. {2.6 d}

In the case of routine, scheduled, or operational delivery/distribution of products, the data service provider provides, including for each product type delivered:

(a) **Product Type Name.** Name of product type. {2.6 a}

(b) **Distribution Destination**. Distinct destinations of operational distribution for type, add to Distribution External Interfaces Count.

(c) **Timeliness.** Timeliness requirement, if any.

(d) **Delivery Means.** Means of delivery (electronic or media, use to sort other items to network or media parameters).

(e) **Delivery Format.** Delivery format, if converted from local production or archive format. (2.6 b}

(f) **Operational Products/Day, Network.** The count of this product type per day delivered operationally by network. {2.6 f}

(g) **Operational Products/Day, Media.** The count of this product type per day delivered operationally by media. {2.6 g}

(h) **Operational Volume/Day, Network.** The volume per day of this product type delivered operationally by network. {2.6 f}

(i) **Operational Volume/Day, Media.** The volume per day of this product type delivered by media. {2.6 g}

Ad hoc, on request delivery or distribution (by network and media) of products the data service provider provides, including:

(a) **Users Requesting Products/Yr.** The number of distinct users requesting products per year.

(b) **User Product Requests/Yr.** The number of product requests received per year.

(c) **By Request Products/Yr., Media.** The number of products provided per year, on media in response to user requests. {2.6 g}

(d) **By Request Products/Yr., Network**. The number of products provided per year, electronically by network. {2.6 f}

(e) **By Request Volume/Yr, Media.** The volume of products provided per year in response to user requests on media. {2.6 g}

(f) **By Request Volume/Yr, Network.** The volume of products provided per year in response to user requests electronically by network. {2.6 f}

(g) **Distribution Format.** Alternative distribution formats offered by a data service provider, where a conversion is done prior to delivery from the locally generated or stored format. {2.6 b}

(h) **Distribution Media Type.** List of types of distribution media used by the data service provider.

(i) **Distribution Media Units/Yr by Type.** The number of units per year of each type of distribution media provided by the data service provider. This can be a forecast capacity for a prospective data service provider. {2.6 h}

(j) **Supporting Data Services.** These services include reformatting, subsetting, packaging, etc. Level of service, values: 1 - supporting services available for most archived data and products; 2 - for less than half of archived data and products; 3 - for a few selected data and products only. {2.6 c}

(k) **Network Distribution Response Time.** The average time from when a product request is received and when it is made available for network delivery, a level of service parameter, values: 1 - ten seconds for software access; 2 - ten seconds for FTP pull/push (or equivalent); 3 - ten minutes; 4 - twenty four hours. {2.6 f}

(l) **Media Distribution Response Time.** The average time from when a product request is received and when it is written to distribution media, packaged, and ready for shipment, a level of service parameter, values: 1 - three days; 2 - one week; 3 - one month. {2.6 g}

(m)**Transmigration Start.** Mission year when migration begins of data, products, and documentation to be transferred another data service provider (e.g. Backbone Data Center or Long Term Archive Center) according to site's Life Cycle Data Management Plan (can be at end of mission, or when products are no longer needed by the site for its mission). {2.6 i}

User Support

These parameters describe or relate to user support provided by the data service provider.

2.2.7.1 Internal Computed Parameters

(a) **Total User Support FTE.** The total estimated annual FTE effort for the User Support functional area, including any effort in addition to the direct user support effort.

(b) **User Support Management FTE.** Includes direct management associated with the User Support functional area. Computed from technical and operations staffing.

(c) **User Support Technical FTE.** Includes technical staff exclusive of direct user support staff.

(d) **User Support Ops FTE**. The estimated annual FTE effort for direct user support and outreach.

(e) **Direct User Support FTE**. The estimated annual FTE effort for direct user support.

(f) **User Contacts/Yr per FTE.** The annual number of user contacts divided by the total effort for user support. Applies to User Support functional area.

(g) **User Contacts/Yr per Ops FTE.** The annual number of user contacts divided by the FTE effort for direct user support. Applies to User Support functional area.

(h) **Outreach FTE.** The estimated annual FTE for outreach effort.

2.2.7.2 User Input Parameters

(a) **User Support Staff Expertise Index.** A general measure or index of the expertise of user support provided by the site, a level of service parameter, values: 1 - science expertise, data structures and tools expertise, format detail expertise, holdings and order/delivery options expertise; 2 - data structures and tools expertise, format detail expertise, holdings and order/delivery options expertise; 3 - format detail expertise, holdings and order/delivery options expertise; 4 - holdings and order/delivery options expertise. {2.7 b}

(b) **Users.** The number of distinct users who contact user support staff in the course of a year. {2.7 a}

(c) **User Support Staffing Target.** The number of user support staff as a function of the user base size, a level of service parameter, values: 1 - one user support staff member per 100 active users; 2 - one per 500; 3 - one per 1,000. {2.7 a}

(d) **Help Desk Hours of Operation.** The hours of operation of a staffed 'help desk' function, a level of service parameter, values: 1 - 7 days/week x 24 hours/day; 2 - five days/week x 12 hours/day; 3 - 5 days/week x 8 hours/day. {2.7 c}

(e) **User Contacts/Yr.** A count of all user contacts - emails, phone calls, etc., handled by the site's user support staff.

(f) **Outreach Activity.** A measure of outreach activities performed by the data service provider, a level of service parameter, values: 1 - training sessions, expanded booth support at four conferences/year, produce and distribute outreach material; 2 - expanded booth support at four conferences/year, produce and distribute outreach material; 3 - booth support at four conferences/year, produce and distribute outreach material; 4 - produce and distribute outreach material. {2.7 e}

Instrument / Mission Operations

These parameters describe or relate to instrument and, if applicable, mission operations functions performed by the data service provider. Instrument monitoring, command generation, event scheduling, etc., is assumed to be a 24x7 activity.

2.2.8.1 Internal Computed Parameters

(a) **Total Instrument FTE.** The total estimated annual FTE effort for the Instrument / Mission Operations functional area, including any effort in addition to actual operational effort.

(b) **Instrument Management FTE.** Includes direct management associated with the Instrument / Mission Operations functional area. Computed from technical and operations staffing.

(c) **Instrument Technical FTE.** Includes technical staff exclusive of direct operations staff.

(d) **Instrument Ops FTE**. The estimated total annual FTE effort for platform and instrument operations.

(e) **Platform Operations FTE.** Includes monitoring status and performance of, and generate commands for spacecraft.

(f) **Instrument Operations FTE.** Includes monitoring status and performance of, and generate commands for, instrument(s).

2.2.8.2 User Input Parameters

(a) **Platforms Monitored.** The number of platforms whose performance, health and safety, etc., are monitored by the data service provider. {2.8 a}

(b) **Platform Actions/Yr.** The number of platform commands generated for upload, platform events scheduled, etc., per year. {2.8 a}

(c) **Platform Flag.** Indicates whether or not the data service provider uses the services of a platform operator's mission operations system (e.g. provides commands to a NASA or other operator facility for validation and uploading). {2.8 b}

(d) **Instruments Monitored.** The number of instruments the data service provider is responsible for monitoring. {2.8 a}

(e) **Instrument Actions/Yr.** The number of instrument commands generated for upload, instrument events scheduled, et., per year. {2.8 a}

Sustaining Engineering

These parameters describe or relate to sustaining engineering (i.e. software maintenance and enhancement of operational software) performed by the data service provider.

2.2.9.1 Internal Computed Parameters

(a) **Total Sustaining Engineering FTE.** The total estimated annual FTE effort for the Sustaining Engineering functional area, including any effort in addition to actual operational effort.

(b) **Sustaining Engineering Management FTE.** Includes direct management associated with the Sustaining Engineering functional area. Computed from technical staffing.

(c) **Sustaining Engineering Technical FTE.** Includes technical staff engaged in software maintenance.

(d) **SLOC Maintained.** The number of lines of code that are maintained by the site, of custom (site developed rather than COTS) software used to support the functional areas. Includes reused software. Maintenance is assumed to be equivalent to sustaining engineering - enhancement as well as bug fixes. {2.9 a}

2.2.9.2 User Input Parameters

Sustaining Engineering LOS. Level of service indicated by allowed impact on operations, values: 1 - no or very infrequent interruptions; 2 - occasional interruptions; 3 - interruptions a secondary concern. {2.9 a}

Engineering Support

These parameters describe or relate to engineering support provided by the data service provider.

2.2.10.1 Internal Computed Parameters

(a) **Total Engineering Support FTE.** The total estimated annual FTE effort for the Engineering Support functional area, including any effort in addition to actual operational effort.

(b) **Engineering Support Management FTE.** Includes direct management associated with the Engineering Support functional area. Computed from technical staffing.

(c) **Engineering Support FTE.** Includes engineering and technical effort that is not otherwise called out, e.g. system engineering, network engineering, test engineering, system administration, and database administration.

2.2.10.2 User Input Parameters

(a) **Technical LOS.** Technical (system administration, network administration, database administration, security, etc.) level of service indicated by allowed impact on operations, values: 1 - no or very infrequent interruptions; 2 - occasional interruptions; 3 - interruptions a secondary concern. {2.10 a}

(b) **Engineering LOS.** Engineering (systems engineering, test engineering, configuration management, etc.) level of service indicated by allowed impact on operations, values: 1 - no or very infrequent interruptions; 2 - occasional interruptions; 3 - interruptions a secondary concern. {2.10 b}

Technical Coordination

These parameters describe or relate to technical coordination performed by the data service provider.

2.2.11.1 Internal Computed Parameters

(a) **Total Technical Coordination FTE.** The total estimated annual FTE effort for the Technical Coordination functional area.

(b) **Technical Coordination Management FTE.** Includes direct management associated with the Technical Coordination functional area. Computed from technical staffing.

(c) **Technical Coordination FTE.** Includes technical staff directly engaged in technical coordination.

(d) **Architecture and IT Coordination FTE.** {2.11 a}

(e) **Data Stewardship Coordination FTE.** {2.11 b}

(f) **Best Practices and Quality Coordination FTE.** {2.11 c}

(g) **Standards and Interfaces Coordination FTE.** {2.11 d}

(h) **Inter-Provider Coordination FTE**. {2.11 e}

(i) **User Support Coordination FTE.** {2.11 f}

(j) **Security Coordination FTE.** {2.11 g}

(k) **Metrics Coordination FTE.** {2.11 h}

2.2.11.2 User Input Parameters

(a) **Architecture and IT Coordination Flag.** Set to 1 if this activity is required, else 0. {2.11 a}

(b) **Data Stewardship Coordination Flag.** Set to 1 if this activity is required, else 0. {2.11 b}

(c) **Best Practices and Quality Coordination Flag.** Set to 1 if this activity is required, else 0. {2.11 c}

(d) **Standards and Interfaces Coordination Flag.** Set to 1 if this activity is required, else 0. {2.11 d}

(e) **Inter-Provider Coordination Flag**. Set to 1 if this activity is required, else 0. {2.11 e}

(f) **User Support Coordination Flag.** Set to 1 if this activity is required, else 0. {2.11 f}

(g) **Security Coordination Flag.** Set to 1 if this activity is required, else 0. {2.11 g}

(h) **Metrics Coordination Flag.** Set to 1 if this activity is required, else 0. {2.11 h}

(i) **Technical Coordination Travel Budget.** Annual budget for travel associated with technical coordination. {2.11 i}

Implementation

These parameters describe or relate to system implementation performed by the data service provider.

2.2.12.1 Internal Computed Parameters

(a) **Total Implementation FTE.** The total annual estimated FTE for the implementation area.

(b) **Implementation Management FTE.** Includes direct management associated with implementation.

(c) **Software Development FTE.** The total estimated annual FTE for data system software development, integration, and test, if this is computed by functional area. This will be projected from the amount of software to be developed and the implementation period. {2.12 d}

(d) **Applications Software Development FTE.** The estimated annual effort for applications software development for user data services, etc., beyond the base data system. {2.12 c}

(e) **Implementation Engineering FTE.** The estimated annual effort for engineering support to system development, e.g. system integration and test, configuration management. {2.12 d}

(f) **Custom Software, SLOC.** The size of the software required, if this is computed by functional area. This will be projected from mission parameters that size the system needed.

2.2.12.2 User Input Parameters

(a) **Software Reuse Fraction.** The amount of software that will be reused from previous projects for base data system development. The precise formulation is TBD; it must allow for rework of reused software, etc.

(b) **Applications Software Development LOS.** A level of service parameter that scopes the applications software development (above the base data system) to meet specific user needs, values: 1 - data mining or data integration, custom science product, data manipulation tools; 2 - custom science product, data manipulation tools; 3 - data manipulation tools; 4 - none. {2.12 e}

2.2.12.3 Cost Model Output Parameters

(a) **Development Staff, FTE.** The annual FTE for development effort, technical excluding management.

(b) **Development Staff Cost.** The annual cost for development staff, using the development staff labor rate.

(c) **Hardware Purchase Cost.** The annual cost for data system hardware needed by the data service provider. This will be projected from mission parameters that size the system needed.

(d) **COTS Software Purchase / License.** The cost for purchase of COTS software package and/or annual license costs.

(e) **Facility Preparation Cost.** All costs associated with preparation of the facility to house the data service provider, and lease, utilities, etc., during the implementation period.

(f) **Total Implementation Period FTE.** The annual sum of all implementation period FTE components.

(g) **Total Implementation Period Cost.** The annual sum of all implementation period cost elements.

Management

These parameters describe or relate to management, administrative, and related functions performed by the data service provider.

2.2.13.1 Internal Computed Parameters

(a) **Total Management FTE.** The total estimated annual FTE effort for the Management functional area.

(b) **Center-Level Management FTE.** Includes center level ‘front office’ management and administration. Computed from overall functional area staffing. {2.13 a}

(c) **Functional Area Management FTE.** Includes the sum of the direct management FTE associated with the other functional areas. Computed from functional area management FTE parameters. {2.13 a}

(d) **Planning and Coordination FTE.** The effort associated with coordinating with other ESE data service providers and ESE in management areas such as strategic planning, policies, etc. {2.13. b}

(e) **Science Coordination FTE.** The effort associated with coordination of science activities, internal and with ESE and other data service providers, peer review, user advisory processes, etc. {2.13.c}

(f) **System Engineering Coordination FTE.** The effort associated with internal system engineering coordination (e.g. planning technology refreshes, etc.). {2.13 d}

(g) **Data Stewardship Coordination FTE.** The effort associated with internal data stewardship, data administration, data management planning, etc. {2.13 e}

2.2.13.2 User Input Parameters

None.

Facility / Infrastructure

These parameters describe or relate to facility support and infrastructure maintenance performed by the data service provider.

2.2.14.1 Internal Computed Parameters

(a) **Total Facility / Infrastructure FTE.** Includes the sum of Facility / Infrastructure elements.

(b) **Facility / Infrastructure Management FTE.** Effort for managing Facility / Infrastructure activities.

(c) **Logistics Support FTE.** Includes property management, logistics, consumables procurement, facility support, etc., within the data service provider. {2.14 d}

(d) **Security FTE.** Includes physical and IT security effort. {2.14 a}

2.2.14.2 User Input Parameters

(a) **External Net Connection**. A list of external network connections that the data service provider supports.

(b) **Source / Service.** The vendor that is the source of the network connection or provider of the network service.

(c) **Bandwidth.** The nominal bandwidth or class of service or capacity of the network connection.

(d) **Recurring COTS Software License Cost.** Cost of annual renewal / update of COTS licenses. Placeholder for now.

(e) **Facility Area**. The area in square feet required to house the data service provider.

(f) **Data System Area.** The area within the facility required to house the data service provider's data system(s).

(g) **Backup Archive Facility LOS.** A level of service parameter characterizing the backup archive facility provided by the site, values: 1 - an environmentally controlled and physically secure off-site backup archive facility; 2 - an on-site but separate environmentally controlled and physically secure off-site backup facility; 3 - a backup capability within the data service provider's primary data system(s). {2.14 c}

(h) **Internal Support LOS.** The level of service associated with resource planning, logistics, facility management, etc., values: 1 - no or very infrequent interruption of operations; 2 - occasional interruptions of operations; 3 - as needed, with interruption of operations a secondary concern. {2.14 d}

2.2.14.3 Cost Model Output Parameters

(a) **Recurring Network / Communications Cost.** The cost associated with network connectivity required by the data service provider.

(b) **Recurring COTS Software Cost.** The cost for COTS upgrades or licenses during the operating period.

(c) **Hardware Maintenance Cost.** The annual cost of maintaining the system hardware, assumed to be TBD a fraction of the hardware purchase cost.

(d) **Supplies Cost.** The annual cost of supplies, including storage and distribution media.

(e) **Recurring Facility Cost.** The total annual facility cost, including lease, utilities, etc., during the operating period.

Site Level Parameters

These parameters describe or relate to the data service provider site as a whole. In some cases they are roll-ups of (selected) functional area parameters listed above.

2.2.15.1 Internal Computed Parameters

None

2.2.15.2 User Input Parameters

None

2.2.15.3 Cost Model Output Parameters

(a) **Management Staff, FTE.** The annual FTE associated with management and administration, including financial administration, supervision, and other administrative functions. Includes overall data service provider management as well as management associated with individual functional areas.

(b) **Management Staff Cost.** The cost for management staff, above, using the management staff labor rate.

(c) **Technical Coordination Staff FTE.** The annual FTE associated with supporting SEEDS technical coordination processes, including developing and maintaining common standards and interfaces.

(d) **Technical Coordination Staff Cost.** The cost for technical coordination staff, above, using the technical coordination staff labor rate.

(e) **Sustaining Engineering FTE.** The annual FTE associated with sustaining engineering, which includes bug fixes and enhancements to custom software.

(f) **Sustaining Engineering Staff Cost.** The cost for sustaining engineering staff, using the sustaining engineering staff labor rate.

(g) **Engineering Support FTE.** The annual FTE associated with system engineering, system administration, database administration and other general technical support.

(h) **Engineering Support Staff Cost.** The cost for engineering support staff, using the engineering support labor rate.

(i) **Operations Staff FTE.** The annual FTE for all aspects of data service provider operations, including system operations, user support, etc.

(j) **Operations Staff Cost.** The cost for operations staff, using the operations staff labor rate.

(k) **Total Operating FTE.** The annual sum of the operating FTE components.

(l) **Total Operating Cost.** The annual sum of all operating cost elements.

Control Parameters

These parameters provide control information for execution of the cost estimation model. Some apply across data service providers, rather than to a particular data service provider.

2.2.16.1 Internal Computed Parameters

None

2.2.16.2 User Input Parameters

(a) **Annual Inflation Rate.** The annual rate of inflation to be applied to all recurring staff costs, lease costs, or license costs.

(b) **Processing Hardware Discount Rate.** The annual rate at which the cost of processing hardware of constant capacity is projected to decline, 50% in 18 months (i.e. capacity per unit cost doubles in 18 months).

(c) **Storage Hardware Discount Rate.** The annual rate at which the cost of storage hardware of constant capacity is projected to decline, 50% in 12 months (i.e. capacity per unit cost doubles in 12 months).

(d) **Network Capacity Discount Rate.** The annual rate at which the cost of constant network capacity or bandwidth is projected to decline, 50% in 9 months (i.e. capacity per unit cost doubles in 9 months).

(e) **COTS Software Discount Rate.** The annual rate at which the cost of COTS software of constant capability is projected to decline. (TBD)

(f) **Implementation Period.** The number of mission years over which development costs are spread - implementation is assumed to start with mission year 1.

(g) **Management Staff Labor Rate**. The fully loaded labor rate for management and administration.

(h) **Technical Coordination Staff Labor Rate.** The fully loaded labor rate for technical coordination.

(i) **Development Staff Labor Rate.** The fully loaded labor rate for development staff.

(j) **Operations Period.** The number of mission years over which operations costs are spread.

(k) **Operations Start.** The mission year during which operations are assumed to start.

(l) **Operations Staff Labor Rate.** The fully loaded labor rate for operations staff.

(m)**Sustaining Engineering Staff Labor Rate.** The fully loaded labor rate for sustaining engineering.

(n) **Engineering Support Labor Rate.** The fully loaded labor rate for engineering support.

Cost Estimation Model Output Parameters

These parameters, defined in section 2.2 above, are the output that will be produced by the cost estimation model; i.e. they comprise the initial draft of the content of the cost estimate. They are grouped into costs (and support information) for the initial implementation period, followed by costs (and support information) for the operations period.

Initial Implementation Period

(a) Management Staff, FTE.

(b) Management Staff Cost.

(c) Technical Coordination Staff FTE.

(d) Technical Coordination Staff Cost.

(e) Development Staff, FTE.

(f) Development Staff Cost.

(g) Hardware Purchase Cost.

(h) COTS Software Purchase / License.

(i) Facility Preparation Cost.

(j) Total Implementation FTE.

(k) Total Implementation Cost.

Operations Period

(a) Management Staff FTE.

(b) Management Staff Cost.

(c) Technical Coordination Staff FTE.

(d) Technical Coordination Staff Cost.
(e) Sustaining Engineering FTE.
(f) Sustaining Engineering Cost.
(g) Engineering Support FTE.
(h) Engineering Support Cost.
(i) Operations Staff FTE.
(j) Operations Staff Cost.
(k) Development Staff FTE.
(l) Development Staff Cost.
(m) Recurring Network / Communications Cost.
(n) Recurring COTS Software Cost.
(o) Hardware Purchase Cost.
(p) Hardware Maintenance Cost.
(q) Supplies Cost.
(r) Recurring facility Cost.
(s) Total Operating FTE.
(t) Total Operating Cost.

Cost Estimation Model User Input Parameters

These parameters, defined in section 2.2 above, must be provided by the user when executing the cost estimation model.

These parameters apply to both implementation and operations. They include control parameters that apply to the data service provider, such as labor rates and planned implementation and operation periods, and parameters that describe the mission workload planned for the data service provider. These mission parameters drive the sizing of the data service provider, and the sizing drives the estimated costs.

Control Parameters

These are overall control parameters that are required for any data service provider whose costs are to be estimated.

(a) Annual Inflation Rate.
(b) Hardware Discount Rate.
(c) COTS Software Discount Rate.
(d) Implementation Period.
(e) Management Staff Labor Rate.

(f) Technical Coordination Staff Labor Rate.

(g) Development Staff Labor Rate.

(h) Operations Period.

(i) Operations Start.

(j) Operations Staff Labor Rate.

(k) Sustaining Engineering Staff Labor Rate.

(l) Engineering Support Labor Rate.

Others TBD.

Mission Parameters

This set of parameters constitutes a complete description of the mission requirements the data service provider must meet, and thus constitutes the sizing information for the data service provider. These parameters are derived from mission descriptions for data service providers. Mission descriptions from actual data service providers will be used to build the comparables database, and mission descriptions for future data service providers will be a source for cost estimation input parameters.

Mission parameters will be listed by functional area in the sections that follow below. Each section will contain a list of the information that will be collected from data service providers for that area. Some of the information is needed for a background understanding of how the data service provider functions and is more directly related to the requirements and levels of service discussed in Section 5.

2.4.2.1 Ingest

Mission parameters for the ingest function are drawn from a description of the data or product streams the data service provider ingests. The description includes the information listed below for each data or product type.

(a) Product Type Name.

(b) Ingest External Interfaces.

(c) Ingest Source.

(d) Ingest Delivery Means.

(e) Ingest LOS for Product Type.

(f) Products of Type Ingested Per Day.

(g) Volume of Type Ingested Per Day.

(h) Ingest Product Type Format.

(i) Conversion Format for Product Type.

(j) Ingest Product Type Retention Period.

2.4.2.2 Processing

Mission parameters for the processing function are drawn from a description of the product streams the data service provider generates. The description includes the information listed below for each data or product type.

(a) Product Type Name.
(b) Product Type Software Source.
(c) Product Type QA Function.
(d) Production Mode for Type.
(e) Operational Production Mode for Type.
(f) Products of Type Generated per Day.
(g) Volume of Type Generated per Day.
(h) Product Type Format.
(i) Generated Product Type Retention Period.
(j) Reprocessing Capacity for Type.
(k) Reprocessing Plan for Type.
(l) Operational Processing LOS by Type.
(m) Non-Operational Processing LOS by Type.
(n) Reprocessing LOS by Type.
(o) Science Software LOS.

2.4.2.3 Documentation

TBD. Mission parameters for the documentation function are drawn from a description of the product streams the data service provider ingests or generates and adds to its archive. The scope of the documentation can be indicated by a) a count of the product types the data service provider handles, since there can be extensive documentation of each product type, and b) a count of the number of product instances the data service provider handles, since there will be documentation associated with each product instance, if only to identify its unique spatial and temporal coverage. Another dimension is the complexity of the documentation, which may be driven by documentation standards that the data service provider uses on its own accord or is required to use.

(a) Documentation LOS.
(b) User Comment LOS.
(c) DIFs Delivered/Yr.

2.4.2.4 Archive

Mission parameters for the processing function are drawn from a description of the product streams the data service provider ingests and generates. Details concerning the retention on the archive of data and products ingested by the data service provider from external sources or generated locally by the data service provider are included in the ingest and processing information described above. The required archive capacity can be projected from that information.

(a) Archive Media Type.
(b) Archive Media Standard.
(c) Archive Media Unit Capacity.
(d) Archive Media Fill Rate.
(e) Archive Capacity.
(f) Archive Backup LOS.
(g) Archive Backup Fraction.
(h) Archive Backup Plan.
(i) Archive Migration Plan.
(j) Archive Monitoring.
(k) Archive Entry/Exit Screening.

2.4.2.5 Search and Order

Mission parameters for the search and order function are drawn from a description of the catalog search and order services the data service provider provides.

(a) Search and Order Scope.
(b) Internal Catalog Search Complexity.
(c) External Catalog Search and Order.
(d) Descriptive Information LOS.
(e) System-System Search.

2.4.2.6 Access and Distribution

Mission parameters for the access and distribution function are drawn from a description of the operational and ad hoc access and distribution services the data service provider provides.

(a) Distribution External Interfaces.
(b) Access and Distribution Scope.
(c) Access and Distribution Service Modes.

Routine, scheduled, or operational delivery/distribution of products the data service provider provides, including for each type delivered:

(a) Product Type Name.
(b) Distribution Destination.
(c) Timeliness.
(d) Delivery Means.
(e) Delivery Format.
(f) Operational Products/Day, Network.
(g) Operational Products/Day, Media.
(h) Operational Volume/Day, Network.
(i) Operational Volume/Day, Media.

Ad hoc, on request delivery or distribution of products the data service provider provides, including:

(a) Users Requesting Products/Yr.
(b) Product Requests Received/Yr.
(c) By Request Products/Yr, Media.
(d) By Request Products/Yr, Network.
(e) By Request Volume/Yr, Media.
(f) By Request Volume/Yr, Network.
(g) Distribution Format.
(h) Distribution Media Type.
(i) Distribution Media Units/Yr by Type.
(j) Supporting Data Services.
(k) Network Distribution Response Time.
(l) Media Distribution Response Time.
(m) Transmigration Start.

2.4.2.7 User Support

User support services provided by the data service provider, including:

(a) User Support Staff Expertise Index.
(b) Users.
(c) User Support Staffing Target.
(d) Help Desk Hours of Operation.
(e) User Contacts Per Year.

(f) Outreach Activity.

2.4.2.8 Instrument / Mission Operations

Instrument monitoring, command generation, event scheduling, etc., is assumed to be a 24x7 activity.

(a) Platforms Monitored.
(b) Platform Actions/Yr.
(c) Platform Flag.
(d) Instruments Monitored.
(e) Instrument Actions/Yr.

2.4.2.9 Sustaining Engineering

(a) Sustaining Engineering LOS.

2.4.2.10 Engineering Support

(a) Technical LOS.
(b) Engineering LOS.

2.4.2.11 Technical Coordination

(a) Architecture and IT Coordination Flag.
(b) Data Stewardship Coordination Flag.
(c) Best Practices and Quality Coordination Flag.
(d) Standards and Interfaces Coordination Flag.
(e) Inter-Provider Coordination Flag.
(f) User Support Coordination Flag.
(g) Security Coordination Flag.
(h) Metrics Coordination Flag.
(i) Technical Coordination Travel Budget.

2.4.2.12 Implementation

(a) Software Reuse Fraction.
(b) Applications Software Development LOS.

2.4.2.13 Management

None.

2.4.2.14 Facility / Infrastructure

These are non-staff items required to support data service provider operations.

(a) External Net Connection.
(b) Source / Service.
(c) Bandwidth.
(d) Recurring COTS Software License Cost.
(e) Facility Area.
(f) Data System Area.
(g) Backup Archive Facility LOS.
(h) Internal Support LOS.

Mapping of Reference Model Parameters to Requirements / Levels of Service

This section presents a mapping of the requirements / levels of service defined in Working Paper 5, "General DSP Reference Model - Requirements / Levels of Service" with the Data Service Provider Reference Model parameters defined in Section 2 above.

In this version, the mapping takes the form of simple tables, one for each functional area. Each table lists the requirements identified by their "WP-5 Requirement ID", the Working Paper 5 subsection number in which they appear, with the applicable parameters identified by their "Parameter ID", which is the abbreviated Section 2.2 subsection number in which they appear and their item number within the subsection (e.g. a parameter ID of 72-4 refers to section 2.2.7.2 item 4, "Help Desk Hours of Operation"; the intent is to make it easy to find the parameter definition). LOS in parentheses following the requirement ID indicates that levels of service were defined for the requirement, and LOS after the parameter ID indicates that it holds the LOS value. The parameters are for convenience grouped as "computed", i.e. internal parameters derived from inputs, or "input", i.e. parameters whose values would be specified by the user.

The cost estimation parameters in Section 2.2 are mapped to requirements in Working Paper 5 in order to ensure that the cost estimate is driven by real requirements. For example, assume that a data service provider will have to ingest a certain volume of data in order to meet its mission responsibilities. The volume of data it must ingest will affect its implementation and operation cost. Therefore in Working Paper 5 there will be a requirement that the data service provider shall ingest a certain volume (or that the data service provider shall ingest a list of data streams whose volume totals to a certain volume), and in Section 2.2 there will be a corresponding operational workload parameter, the volume of data to be ingested, that is needed for cost estimation.

The parameters that map most directly to the requirements will be included in the tables below, along with some computed 'roll-up' parameters (e.g. 'product types ingested/yr' is included as well as the parameter holding the list of types, 'ingest product type name'); other computed parameters, derived from listed parameters, will not be included. Because of the general nature of the requirements in Working Paper 5, there will often be a 'many to one' mapping of parameters to requirements.

Table 1 - Ingest Requirements/LOS vs Parameters

WP-5 Requirement ID	Parameter ID's and Notes
2.1 a (LOS)	Input: 12-1, 12-3, 12-5-LOS, 12-6, 12-7, 12-8, 12-9, 12-10 Computed: 11-5, 11-8, 11-9, 11-10, 11-13-LOS
2.1 b	TBD - metrics collection

Table 2 - Processing Requirements/LOS vs Parameters

WP-5 Requirement ID	Parameter ID's and Notes
2.2 a (LOS)	Input: 22-1, 22-4, 22-5, 22-6, 22-7, 22-8, 22-9, 22-12-LOS Computed: 21-5, 21-7, 21-9, 21-12, 21-14, 21-16, 21-18, 21-21-LOS
2.2 b (LOS)	Input: 22-1, 22-4, 22-6, 22-7, 22-8, 22-9, 22-14-LOS Computed: 21-6, 21-7, 21-9, 21-13, 21-14, 21-16, 21-18, 21-22-LOS
2.2 c (LOS)	Input: 22-1, 22-6, 22-7, 22-10-LOS Computed: 21-8, 21-9, 21-15, 21-16, 21-18, 21-23
2.2 d (LOS)	Input: 22-1, 22-6, 22-7, 22-11, 22-14-LOS Computed: 21-8, 21-9, 21-15, 21-16, 21-18

2.2 e (LOS)	Input: 22-2, 22-15-LOS Computed: 21-3
2.2 f	Input: 22-16 Computed: 21-3
2.2 g	TBD - metrics collection

Table 3 - Documentation Requirements/LOS vs Parameters

WP-5 Requirement ID	Parameter ID's and Notes
2.3 a (LOS)	Input: 32-1-LOS
2.3 b (LOS)	Input: 32-2-LOS
2.3 c	Input: 32-3

Table 4 - Archive Requirements/LOS vs Parameters

WP-5 Requirement ID	Parameter ID's and Notes
2.4 a	Input: 12-10, 22-9 Computed: 41-5, 41-8, 41-10, 41-13, 41-15
2.4 b	Input: 12-10 Computed: 41-5, 41-8, 41-10, 41-13, 41-15
2.4 c	Input: 12-10, 22-9, 42-8, 42-10-LOS, 42-11-LOS
2.4 d	None.
2.4 e (LOS)	Input: 42-5-LOS
2.4 f (LOS)	Input: 42-11-LOS
2.4 g (LOS)	Input: 42-10-LOS
2.4 h (LOS)	Input: 42-6-LOS, 42-7, 42-8-LOS Computed: 41-14, 41-15
2.4 i (LOS)	Input: 42-2-LOS
2.4 j (LOS)	Input: 42-9-LOS
2.4 k	TBD Metrics Collection

Table 5 - Search and Order Requirements/LOS vs Parameters

WP-5 Requirement ID	Parameter ID's and Notes
2.5 a (LOS)	Input: 52-1-LOS
2.5 b (LOS)	Input: 52-2-LOS
2.5 c (LOS)	Input: 52-4-LOS
2.5 d	Input: 52-5
2.5 e (LOS)	Input: 52-3-LOS

Table 6 - Access and Distribution Requirements/LOS vs Parameters

WP-5 Requirement ID	Parameter ID's and Notes
2.6 a (LOS)	Input: 62-2-LOS
2.6 b	Input: 62-8, 62-19 Computed: 61-9
2.6 c (LOS)	Input: 62-22-LOS
2.6 d	Input: 62-3
2.6 e	Input: 62-23-LOS
2.6 f (LOS)	Input: 62-9, 62-11, 62-16, 62-18, 62-23-LOS Computed: 61-20, 61-21
2.6 g (LOS)	Input: 62-10, 62-12, 62-15, 62-17, 62-24-LOS Computed: 61-22, 61-23
2.6 h	Input: 62-21 Computed: 61-24, 61-25
2.6 i	Input: 62-25 Computed: 61-26, 61-27
2.6 j	TBD - Metrics Collection

Table 7 - User Support Requirements/LOS vs Parameters

Requirement ID	Parameter ID's and Notes
2.7 a (LOS)	Input: 72-3-LOS
2.7 b (LOS)	Input: 72-1-LOS

2.7 c (LOS)	Input: 72-4-LOS
2.7 d	None.
2.7 e (LOS)	Input: 72-6-LOS

Table 8 - Instrument / Mission Operations Requirements/LOS vs Parameters

Requirement ID	Parameter ID’s and Notes
2.8 a	Input: 82-1, 82-2, 82-4, 82-5
2.8 b	Input: 82-3

Table 9 - Sustaining Engineering Requirements/LOS vs Parameters

Requirement ID	Parameter ID’s and Notes
2.9 a (LOS)	Input: 92-2-LOS Computed: 91-4

Table 10 - Engineering Support Requirements/LOS vs Parameters

Requirement ID	Parameter ID’s and Notes
2.10 a (LOS)	Input: 102-1-LOS
2.10 b (LOS)	Input: 102-2-LOS

Table 11 - Technical Coordination Requirements/LOS vs Parameters

Requirement ID	Parameter ID’s and Notes
2.11 a	Input: 112-1 Computed: 111-4
2.11 b	Input: 112-2 Computed: 111-5
2.11 c	Input: 112-3 Computed: 111-6
2.11 d	Input: 112-4 Computed: 111-7

2.11 e	Input: 112-5 Computed: 111-8
2.11 f	Input: 112-6 Computed: 111-9
2.11 g	Input: 112-7 Computed: 111-10
2.11 h	Input: 112-8 Computed: 111-11
2.11 i	Input: 112-9

Table 12 - Implementation Requirements/LOS vs Parameters

Requirement ID	Parameter ID’s and Notes
2.12 a	None - system design
2.12 b	None - staffing plan
2.12 c	None - facility plan
2.12 d	Computed: 121-3, 121-5
2.12 e (LOS)	Input: 122-2
2.12 f	None - staff

Table 13 - Management Requirements/LOS vs Parameters

Requirement ID	Parameter ID’s and Notes
2.13 a	Computed: 131-2, 131-3
2.13 b	Computed: 131-4
2.13 c	Computed: 131-5
2.13 d	Computed: 131-6
2.13 e	Computed: 131-7

Table 14 - Facility / Infrastructure Requirements/LOS vs Parameters

Requirement ID	Parameter ID’s and Notes
2.14 a	Computed: 141-4
2.14 b	Input: 142-5, 142-6
2.14 c (LOS)	Input: 142-7-LOS
2.14 d (LOS)	Input: 142-8 Computed: 141-3
2.14 e	Input: 142-1, 142-2, 142-3

References and Acronym List

The References Section and the Acronym List for all of these Working Papers is in the document

“References and Acronyms for the Levels of Service / Cost Estimation Working Papers ”.

SEEDS

Working Paper Five:

Data Service Provider Model, Requirements and Levels of Service

June 24, 2003

G. Hunolt, SGT, Inc.

Outline

Introduction

This working paper is the fifth of a set of papers that describes the SEEDS (Strategic Evolution of Earth Science Enterprise Data Systems) Levels of Service (LOS) / Cost Estimation (LOS/CE) study. The study goal is to develop a cost estimation model and coupled requirements and levels of services to support the SEEDS Formulation team in estimating the life cycle costs of future ESE data service providers and supporting systems, where 'data service provider' is used as a generic term for any data/information related activity. The set of working papers is intended to serve as a vehicle for coordinating work on the study, obtaining feedback and guidance from ESDIS SOO and the user community, and as embryos of reports that will be produced as the task proceeds.

As working papers, each version of each paper that appears represents a snapshot in time, with the work in various stages of completion; as work progresses the content (and sometimes the organization) of the working papers will change reflecting progress made, responses to feedback and guidance received, etc.

This fifth working paper of the set describes the requirements and levels of services component of the general data service provider reference model developed for the LOS/CE study, and will reflects results of the 2002 *and 2003 SEEDS Community Workshops and comments received on the draft SEEDS Formulation Team Recommendations Report*.

Changes made for the June, 2003 version of this working paper are shown in italics.

Levels of Service

A major objective of the LOS/CE study is to assist the SEEDS Formulation Team in establishing the minimum (and recommended) levels of service for ESE data service providers. These levels of service will be refined in a bottoms-up manner through community workshops of potential providers and users of ESE data services. To facilitate this process, a user-oriented view of the levels of service is included in this paper.

Levels of service will be associated with functional requirements, describing different degrees of performance with which the requirement would be met. For example, a functional requirement might be: "The data service provider shall distribute data and products to users on media". Accompanying this requirement might be descriptions of quantitatively distinct levels of service, such as "delivery on media shall be provided within one working day of receipt of a data request", "delivery on media shall be provided within two calendar weeks of receipt of a data request", and "delivery on media shall be provided within one calendar month of receipt of a data request". Which level of service would be most appropriate ('recommended') or acceptable (‘minimum’) for a particular ESE data service provider would depend on its particular mission and the needs of its users. Not all requirements have levels of service associated with them; by their nature, some requirements are either met or not met without any shades of gray.

The levels of service and their associated requirements will also feed into the life cycle cost estimation phase of the study because data service provider costs must be driven by / associated with the levels of service required of the data service provider. Successful development of a life cycle cost estimation capability will be dependent on an accurate assessment of the levels of services needed from ESE data service providers.

The requirements developed by this study are not intended to serve as the complete definition of the requirements side of a contract between an ESE program office and ESE data service providers, or to serve as a basis for procurements. This study ignores questions about what level of requirements will be ‘owned’ at the program level vs by data service providers themselves.

Levels of Service and the Data Service Provider Reference Model

The requirements / levels of service are one of three related aspects of the Data Service Provider Reference Model, a general functional model of a generic data service provider:

1) A set of ‘functional areas’ that collectively comprise the full range of functions that a data service provider might perform and the areas of cost that must be considered by the cost estimation by analogy model. The functional areas of the reference model are defined in Working Paper 3, “Data Service Provider Reference Model - Functional Areas”.

2) A set of parameters for each functional area that constitute a quantitative description of the workload, staff effort, and any other factors that contribute to cost for that area, additional ‘roll-up’ parameters that sum items such as staff effort across the functional areas, and other parameters like labor rates that are required for cost estimation. The parameters of the reference model are defined in detail in Working Paper 4, “Data Service Provider Reference Model - Model Parameters”.

3) A set of requirements and levels of service for each functional area. These are defined in Section 2 of this working paper.

These three aspects of the model are closely coupled to ensure the internal consistency of the model. The set of functional areas is the underpinning; both the model parameters and requirements / levels of service are organized according to the functional areas. The requirements / levels of service and the model parameters are coupled in that the definitions of the requirements / levels of service embody model parameters. This integration of the three aspects of the model is intended to ensure that estimated costs are driven by and traceable to requirements to the fullest extent possible. Working Paper 4 includes a mapping of the Data Service Provider Reference Model parameters to the requirements / levels of service. The intent is to show which parameters fall within the scope of each requirement, and to ensure that each requirement / levels of service that should have one or more parameters associated with it actually does.

The intent of the requirements / levels of service described in Section 2 below (and the corresponding functional areas presented in Working Paper 3) is to provide a complete description at a reasonable level of detail of the abstract ESE data service provider, and to reflect the concerns expressed in the *2002 and 2003* SEEDS community workshop. The ability of the cost estimation by analogy approach to reflect the full range of detail described in the functional areas and requirements / levels of service will be limited by the information available in the comparables database and the feasibility of reasonable assumptions where information is not available. This will be reflected in the reference model's parameter set, described in Working Paper 4.

As the needs of the ESE science and applications program evolve, and hence the ESE roles and missions for data service providers evolve, and as information technology that touches all aspects of every data service provider and the user community evolves (e.g. the GRID distributed computing approach), the data service provider requirements and levels of service will evolve. The content of this paper can only represent a snapshot in time - and indeed a snapshot that is in part tied to current and recent past experience with working data service providers. If the cost estimation tool (and the underling data service provider model) proves useful, it will have to be maintained and revised perhaps dramatically to preserve or improve its usefulness over time.

In addition to evolving with changing ESE program needs, the cost estimation by analogy model (and the data service provider model) will be improved in successive iterations as the comparables database grows and includes more new activities, and with lessons learned derived from use of earlier versions of the model.

Section 2 presents the requirements / levels of service organized by the reference model's functional areas. Section 3 presents a user-oriented view of the levels of service (with requirements implied rather than stated explicitly).

Appendix A includes a draft set of program level requirements, "NewDISS Level 0 Requirements", GSFC, September 2001, which were used as a starting point / umbrella for the requirements / levels of service in Section 3.

Data Service Provider Requirements / Levels of Service

This section presents the requirements / levels of service template for a generic ESE data service provider, organized by the data service provider reference model's fourteen functional areas. As such it does not imply or embody any architecture, i.e. any allocation of requirements to various particular components.

The term 'template' is used for two reasons. The first is that all of the requirements / levels of service will not apply to all actual ESE data service providers. The second is that the requirements contain placeholders for specifics that must be filled in (i.e. choices between alternatives shown, or between possible levels of service, or replacement of placeholders with lists or numerical values) to generate from the template a set of requirements / levels of service that would apply to a specific ESE data service provider, and that would allow a cost estimate for it to be produced.

Appendix A contains a draft set of high level or programmatic requirements referred to as "NewDISS Level 0 Requirements" produced by the SEEDS Formulation Team in September, 2001. These provide an "umbrella" for the requirements described below. Additional guidance for the initial set of requirements and levels of service was drawn from the ESDIS Project Level 2 Requirements for EOSDIS Version 0, updated March 2000, which addressed requirements and levels of service, the report "Global Change Science Requirements for Long-Term Archiving", NOAA-NASA and USGCRP Program Office, March 1999, and the report "Ensuring the Climate Record from the NPP and NPOESS Meteorological Satellites", NRC Committee on Earth Studies, September 2000.

The requirements and levels of service were updated following the February, 2002, SEEDS community workshop, responding to comments and recommendations made at the workshop and in white papers submitted to the workshop. *The requirements and levels of service were updated again in June, 2003, responding to comments made on the draft SEEDS Formulation Team Recommendations Report and at the March, 2003, SEEDS community workshop.*

Placeholders for items to be specified when the template is to be used to generate requirements for a specific data service provider are enclosed in brackets […].

Ingest Requirements / Levels of Service

The data service provider shall ingest the following data [ingest data stream table, listing for each data stream: name, source, product types ingested, product type format (input and conversion after ingest if any) products ingested per day of each type, volume ingested per day]. The input data streams should cover all data to be received by the center, e.g. satellite data streams, ancillary data products, processed products generated by other data service providers, etc., based on its ESE mission, and accompanying metadata, documentation, retention plan (e.g. a part of a life cycle data management plan) etc.

Levels of Service:

1) operational (time-critical) ingest with immediate verification of data integrity and quality;
2) routine ingest and verification of data quality and integrity without tight time constraints;
3) ad hoc or intermittent ingest on a non-operational basis with verification of data quality and integrity;
4) ad hoc or intermittent ingest on a non-operational basis.

- Levels of service can be mixed within a data service provider; i.e. different levels may be appropriate for different data streams.
- The data service provider shall provide standard metrics on ingest to the SEEDS Office.

Processing Requirements / Levels of Service

a. The data service provider shall generate the following products ('standard products' characterized by a peer reviewed, validated, reasonably stable, 'science quality' processing algorithm), included required Level 1B products [standard product table, listing for each product type/series: name, format, retention plan, product instances produced per day, volume per day, required input data streams] on a highly reliable, operational basis, either on a routine schedule or on-demand, based on its ESE mission.

Levels of Service:

1) operational products shall be generated within 2 days of ingest/availability of required inputs;

2) operational products shall be generated within 7 days of ingest/availability of required inputs;

3) operational products shall be generated within 30 days of ingest/availability of required inputs.

b. The data service provider shall generate the following products [product table, listing for each product type/series: name, format, retention plan, average product instances produced per day, average volume per day, required input data streams] on an ad hoc, non-operational basis. (The product table can refer to known or expected products, or can be used to establish a capacity to support a level of ad hoc product generation (perhaps data mining or data integration) that will be used to support user needs as they arise.)

Levels of Service:

1) specific targets for processing adopted on a case by case basis;

2) general goals for processing;

3) no goals, purely ad hoc processing.

c. The data service provider shall provide near real-time processing of selected product types [standard product table, listing for each product type the applicable level of service].

1) near real-time products shall be generated within 30 minutes of ingest/availability of required inputs;

2) near real-time products shall be generated within 3 hours of ingest/availability of required inputs;

3) near real-time products shall be generated within 24 hours of ingest/availability of required inputs.

d. The data service provider shall provide a capacity for reprocessing of standard products [standard product table] on an ad hoc basis in response to reprocessing requests.

Levels of Service

1) the aggregate capacity for reprocessing shall be 9 times the original aggregate processing rate;

2) the aggregate capacity for reprocessing shall be 6 times the original aggregate processing rate;

3) the aggregate capacity for reprocessing shall be 3 times the original aggregate processing rate.

e. The data service provider shall reprocess standard products [standard product table, listing for each product a reprocessing interval] according to a reprocessing schedule.

Levels of Service:

1) reprocessing shall be performed according to a negotiated reprocessing schedule;

2) reprocessing shall be performed to meet the general goals of a nominal schedule;

3) reprocessing shall be performed following a nominal schedule on a resource / time available basis.

f. The data service provider shall accept science algorithm software from users for [product list], and perform integration and test of the software, and operational execution of the software to produce products.

Levels of Service:

1) the data service provider shall accept standard, research product generation software, and/or data integration and data mining software from users;

2) the data service provider shall accept research product generation software and/or data integration and data mining software from users;
3) the data service provider shall accept standard and/or research product generation software from users;

4) the data service provider shall accept research product generation software from users;

5) the data service provider shall accept standard product generation software from users.

g. The data services provider shall be capable of cross-calibration of data from multiple sources to produce consistent product time series spanning multiple instruments / platforms.

h. The data service provider shall provide standard metrics on production to the SEEDS Office.

Documentation Requirements / Levels of Service

a. The data service provider shall generate and provide ESE/ SEEDS adopted standards compliant catalog information (metadata, including browse) and documentation describing all data and information produced and/or acquired and held by the data service provider, *including complete documentation of all product generation software used by the data service provider*.

Levels of Service:

1) data and product holdings (including multiple versions of products and corresponding documentation as needed) documented to the ESE / SEEDS adopted standard for long term archiving, including details of processing algorithms, processing history, many etc.;

2) documentation ensured to be sufficient for current use (e.g. product type descriptions, product instance (a.k.a. granule) descriptions including version information, FAQs, 'readme's, web pages with links to metadata, user guides, references to journal articles describing the production or use of the data or product);

3) documentation only as received from product provider.

b. The data service provider shall update documentation of data and products with user comments, e.g. on parameter accuracy, product usability, data services available or needed for a product, etc.

Levels of Service:

1) data and products routinely updated with user comments;

2) data and products occasionally updated with user comments;

3) data and products rarely updated with user comments.

c. The data service provider shall generate and provide DIF (Directory Interchange Format) documents to the Global Change Master Directory on all products available from the data service provider prior to their release for distribution.

Archive Requirements / Levels of Service

a. The data service provider shall add to its archive or working storage the following data and products [archive product table, drawn from ingest data stream table, standard product, and ad hoc product tables and reprocessing volume] and related documentation / metadata, *including format libraries used to read and write the data and products.*

b. The data service provider shall provide for secure, permanent storage of data at the "raw" sensor level (NASA Level 0 plus appended calibration and geolocation information).

c. The data service provider shall provide for secure storage of all standard or other science products it produces, *including their documentation,* until the end of the science mission or until transfer to an approved permanent archive, per the applicable life-cycle data management plan (or separate retention plans).

d. The data service provider shall have the capability to selectively replace archived product instances (single or large sets) with new versions, and to selectively update metadata and documentation (e.g. to update quality flags when a product is validated).

e. The data service provider shall provide for an [archive] [working storage] capacity of [number] TB.

Levels of Service:

1) archive capacity is cumulative sum of all data ingested plus all products generated (including allowance for retaining multiple versions of the same product as required to provide needed support to the provider's science or applications community);

2) archive capacity is limited to a specified threshold.

f. The data service provider shall perform quality screening on data entering the archive (e.g. read after write check when data is written to archive media) and exiting the archive (e.g. tracking of read failures and corrected errors or other indication of media degradation on all reads from archive media).

Levels of service:

1) exit and entry screening;

2) entry screening.

g. The data service provider shall take steps to ensure the preservation of data in its archive.

Levels of service:

1) 10% per year random screening to detect and replace failing / degrading media;

2) 5% per year random screening;

3) 1% per year random screening.

h. The data service provider shall provide a backup and restore capability for its [archive] [working storage].

Levels of service:

1) full off-site backup, with regular sampling and exercise of restore capability to verify integrity;

2) partial, [Backup Fraction - % of archive backed up], off-site backup, with sampling;

3) partial, [Backup Fraction - % of archive backed up], on-site backup, with sampling.

i. The data service provider shall use robust archive media.

Levels of Service:

1) archive media consistent with best commercial practice;

2) archive media and system vendor independent;

3) archive media vendor independent.

j. The data service provider shall plan and perform periodic migration of archive to new archive media / technology.

Levels of Service:

1) planned and budgeted for migration;

2) no planned migration, but ad hoc migration as need is seen to arise.

(Note - this requirement would not apply to a data service provider with a shorter lifetime than a migration cycle appropriate for its archive media / technology.)

k. The data service provider shall provide standard metrics on archive to the SEEDS Office.

Search and Order Requirements / Levels of Service

a. The data service provider shall provide users with access to all metadata and information holdings.

Levels of Service:

1) public access to all users;

2) access to the science and applications community, *with at least a minimal capability for public access*;

3) access to a limited team of scientists or applications specialists, *with at least a minimal capability for access to the science and applications community and the public*.

Public access may be limited in some cases by constraints levied by the original data source. Data service providers with a highly focused primary mission (such as a flight project instrument team) may meet the requirement for public access by teaming with other data service providers actively engaged in public access and services.

b. The data service provider shall provide a world wide web accessible search and order capability to [all users (including the general public) consistent with SEEDS standards and practices] [a limited set of science team members]. (Scope consistent with the level of service for requirement 2.5 a above.)

Levels of Service:

1) allow search for instances of multiple product types that pertain to a specified object or phenomenon (e.g. a named hurricane, a volcanic eruption, a field campaign, etc.)

2) allow search for instances of multiple product types by geophysical parameter(s), time, and space *(by named spatial object from a catalog as well as by coordinates)* applied across multiple product types;

3) allow search for instances of multiple product types by common time and space criteria (coincident search);

4) allow search for instances of single product type by time and space criteria;

6) allow search for particular instances of a product type from a list of those available.

c. The data service provider shall provide the user with the option of quickly viewing information describing any product returned as meeting search criteria.

Levels of Service:

1) descriptive information includes detailed algorithm and use explanations, references to a few published papers that describe the production or use of the product, standard guide and DIF metadata.

2) descriptive information includes references to a few published papers that describe the production or use of the product, standard guide and DIF metadata.

3) descriptive information includes standard guide and DIF metadata.

d. The data service provider shall provide an interface for system-system search and order access as well as an interface for human users.

e. The data service provider shall provide an interface to and support selected external catalog search capabilities (e.g. EDG, Mercury, Echo).

Access and Distribution Requirements / Levels of Service

a. The data service provider shall provide users with access to all data, product, *and documentation (including read software)* holdings, including all standard science products (Level 1b, Level 2, and Level 3) produced by the data service provider.

Levels of Service:

1) public access to all users;

2) access to the science and applications community, *with at least a minimal capability for public access*;

3) access to a limited team of scientists or applications specialists, *with at least a minimal capability for access to the science and applications community and the public*.

Public access may be limited in some cases by constraints levied by the original data source. Data service providers with a highly focused primary mission (such as a flight project instrument team) may meet the requirement for public access by teaming with other data service providers actively engaged in public access and services.

Levels of Service:

1) supporting data services available for most archived data and products;

2) supporting data services available for less than half of archived data and products;

3) supporting data services available for a few selected data and products only.

The particular supporting services available would vary on a product by product basis, depending on the nature of the product and the needs of the user community.

b. The data service provider shall provide data to users on an [operational, subscription (i.e. standing order), and/or in response to request] basis. (An operational basis means in part that a data service provider will formally commit in a level of service agreement or equivalent to terms of service.)

c. The data service provider shall provide an interface for system to system network delivery of data and products.

d. The data service provider shall perform timely distribution of data and products to users by network, providing an average distribution volume capacity of [number] TB per day.

Levels of service:

1) availability of a single product for access by user software within ten seconds;

2) availability of a single product for network delivery (e.g. FTP pickup or push) within ten seconds;

3) availability of a single product for network delivery within ten minutes;

4) availability of a single product for network delivery within twenty four hours.

e. The data service provider shall perform timely distribution of data and products to users on SEEDS standard media types in response to user requests, providing an average volume capacity of [number] TB per day.

Levels of Service:

1) shipping of media product within three days of receipt of request;

2) shipping of media product within one week of receipt of request,

3) shipping of media product within one month of receipt of request.

f. The data service provider shall have the capacity to distribute products on an average of [number] media units per day.

g. The data service provider with final ESE archive responsibility (i.e., a Backbone Data Center unless, for example, or a Science Data Service Provider which has held its products to the time for their transfer to the long term archive) shall transfer its data, products, and documentation (done to the long term archive standard) to the designated long term archive according to its Life Cycle Data Management Plan.

h. The data service provider shall provide SEEDS standard metrics on distribution to the SEEDS Office.

User Support Requirements / Levels of Service

a. The data service provider shall be capable of supporting [number] of distinct, active users per year who request and/or access and use data service provider products.

1) one user support staff member per *500* active users;

2) one user support staff member per *1,000* active users;

3) one user support staff member per *2,500* active users.

(The number of active users is the number of distinct users who request, or through an automated means obtain, delivery of data and/or information products per year, *and who might need help in making choices, interpreting formats, etc., as opposed to more "casual" users who might simply access a data service provider website and download a prepackaged product without interacting with the data service provider staff or causing execution of any system process more complicated than a simple download*.)

b. The data service provider shall provide a trained user support staff.

Levels of service:

1) below plus science expertise in data / product quality and their research uses.

2) below plus technical expertise in data structures, use of tools for format conversions, subsetting, analysis, etc.

3) below plus comprehensive knowledge of details of formats for most if not all products;

4) user support staff are knowledgeable about the data service provider's holdings and ordering/delivery options.

(Not all members of the user support staff would necessarily have the highest level of expertise.)

c. The data service provider shall provide a help desk function (i.e., staff awaiting user contacts who can assist in ordering, track and status pending requests, resolve problems, etc.).

Levels of Service:

1) Help desk staffed seven days per week, twenty-four hours per day.

2) Help desk staffed five days per week, twelve hours per day;

3) Help desk staffed five days per week, eight hours per day;

d. The data service provider shall provide on-line user support (FAQ, data, product and service descriptions, etc.).

e. The data service provider shall perform user outreach, education, and training.

Levels of Service:

1) Below plus provide user training sessions at universities, schools, etc.

2) Below plus expanded booth support including mini-workshops, user training sessions;

3) Below plus booth support at four conferences per year;

4) Produce and make available outreach material - pamphlets, brochures, posters, etc.

Instrument / Mission Operations Requirements / Levels of Service

a. The data service provider shall monitor the status and performance of [name] instruments and in some cases also [name] spacecraft for which it is responsible, generating instrument commands and in some cases spacecraft commands as needed.

b. The data service provider shall obtain the services of a NASA (or other spacecraft operator as appropriate) mission operations facility to provide instrument and spacecraft data and to receive, validate, and transmit instrument and/or spacecraft commands to the spacecraft.

Sustaining Engineering Requirements / Levels of Service

a. The data service provider shall maintain and, as needed, enhance custom software it develops to meet its mission needs, and reused software it customizes and integrates, a total of [number] SLOC.

Levels of Service:

1) no or very infrequent interruptions of data service provider operations;

2) occasional interruptions in data service provider operations;

3) as needed, with interruptions in data service provider operations a secondary concern.

Engineering Support Requirements / Levels of Service

b. The data service provider shall perform system administration, network administration, database administration, coordination of hardware maintenance by vendors, and other technical functions as required for performance of its mission.

Levels of Service:

1) no or very infrequent interruptions of data service provider operations;

2) occasional interruptions in data service provider operations;

3) as needed, with interruptions in data service provider operations a secondary concern.

c. The data service provider shall perform systems engineering, test engineering, configuration management, COTS procurement, installation of COTS upgrades, network/communications engineering and other engineering functions as required for performance of its mission.

Levels of Service:

1) no or very infrequent interruptions of data service provider operations;

2) occasional interruptions in data service provider operations;

3) as needed, with interruptions in data service provider operations a secondary concern.

Technical Coordination Requirements / Levels of Service

a. The data service provider shall provide staff required for participation in SEEDS processes, including ESE data services architecture refinement and evolution, and information technology planning.

b. The data service provider shall provide staff required for participation in SEEDS processes to coordinate data stewardship standards and practices and development and maintenance of standards for content of life cycle data management plans.

c. The data service provider shall provide staff required for participation in SEEDS processes to coordinate best practices among ESE data service providers, including quality assurance standards and practices for all phases of data services provider functions.

d. The data service provider shall provide staff required for participation in SEEDS processes, and cooperating with other ESE data service providers in representing ESE / SEEDS in broader community processes, for developing and maintaining common standards and interface definitions, including those that enable interoperability within the ESE / SEEDS environment and with other systems and networks as needed to support the ESE program.

e. The data services provider shall participate in SEEDS level and/or bilateral processes to coordinate production and delivery of products between ESE data service providers.

f. The data services provider shall participate in SEEDS processes for coordinating user support guidelines and practices among ESE data services providers.

g. The data services provider shall provide staff required for SEEDS coordination of security standards and practices to meet NASA or other established security requirements.

h. The data service provider shall provide staff to coordinate standards for common metrics.

i. The data service provider shall provide funding for travel to support technical coordination activities.

Implementation Requirements / Levels of Service

a. The data service provider shall design and a data and information system capable of meeting its mission requirements. The design shall address hardware configuration and interfaces and allocation of function to platform. The design shall address software configuration, including COTS, software re-use, and new custom software to be developed, including science software embodying product generation algorithms and/or software facilitating integration of science software provided by outside source(s).

b. The data service provider shall develop a staffing plan that addresses staff required to implement and operate the data service provider over its planned lifetime. The staffing plan shall include a breakdown of positions and skill levels assigned to functions.

c. The data service provider shall develop a facility plan, including planning for space, utilities, furnishings, etc., required to support its staff, data and information system, data storage, etc., and the environmental conditioning to be provided.

d. The data service provider shall accomplish the implementation of its data and information system, including purchase and installation of hardware, purchase or licensing and installation and configuration of COTS software, modification, installation and configuration of re-use software, development of new custom software, and integration of all components into a tested system capable of meeting the data service provider's mission requirements.

e. The data service provider shall perform ongoing applications software development.

Levels of Service:

1) Below plus implementation of applications software to perform a 'data mining' or data integration operation to meet a user need.

2) Below plus implementation of product generation software embodying science algorithms, e.g. to produce a product to meet a particular user need;

3) Implementation of software tools for use by users to unpack, subset, or otherwise manipulate products provided by the data service provider;

f. The data service provider shall provide the staff needed to accomplish all needed in-house development and test activities.

Management Requirements / Levels of Service

a. The data service provider shall provide management and administrative staff to perform supervisory, financial administration, and other administrative functions.

b. The data service provider shall provide staff required for participation in SEEDS management processes, strategic planning, coordination with other data centers and activities beyond ESE/SEEDS.

c. The data service provider shall provide staff with science expertise to coordinate the science activities within the data service provider and its interaction with the ESE and broader science community, including a visiting scientist program (or equivalent), collaboration among ESE data service providers to support science needs, annual Enterprise peer review, and support for its User Advisory Group and any other advisory activities appropriate given its ESE role and user community.

d. The data service provider shall provide staff with system engineering expertise to plan information technology upgrades / technology refreshes, based on assessments of changing mission or user needs and availability of new technology. (Coordination with other ESE data service providers is included in technical coordination).

e. The data service provider shall provide staff with data management expertise to develop data stewardship practices, perform data administration with science advice (via the User Advisory Group and other appropriate bodies), develop and maintain life cycle data management plans including data migrations. (Coordination with other ESE data service providers is included in technical coordination).

Facility / Infrastructure Requirements / Levels of Service

a. The data service provider shall maintain site, system, and data security according to established NASA or other policies and practices while providing easiest possible access (consistent with required security) to its data and information services for its user community.

b. The data service provider shall provide and maintain a fully furnished and equipped, environmentally controlled, physically secure facility to house its staff, systems, and data and information holdings.

c. The data service provider shall provide a backup facility for its data and information holdings.

Levels of Service:

1) an environmentally controlled and physically secure off-site backup archive facility;

2) an on-site but separate environmentally controlled and physically secure off-site backup facility;

3) a backup capability within the data service provider's primary data system(s).

d. The data service provider shall perform resource planning, logistics, supplies inventory and acquisition, and facility management.

Levels of Service:

1) no or very infrequent interruptions of data service provider operations;

2) occasional interruptions in data service provider operations;

3) as needed, with interruptions in data service provider operations a secondary concern.

e. The data service provider shall provide network connections and services as needed to support its operations.

User-Oriented View of Levels of Service

This section presents a user-oriented view of the requirements / levels of service given in Section 2. The intent is to provide users with a description of data service provider services and levels of service that they would actually see or interact with. The goal is to provide users with a description of services in terms of their experience, to facilitate their critique and review.

Users in this context include other data service providers (such as applications centers who need ESE products as input, or flight projects who provide data to a data service provider for archive and distribution and so interact with the provider's ingest service) as well as 'end users' such as research scientists or applications specialists who interact with the provider's search and order and access and distribution services.

Because the focus of this section is on the user's interaction with the data service provider, services that the user does not directly interact with, such as sustaining engineering or facility support, are not included, even though successful performance of these services is essential to the success of the services the user actually sees. Services that the user actually sees or interacts with are ingest, processing, documentation, search and order, access and distribution, and user support.

The user-oriented service descriptions are organized according to the reference model's functional areas. Each service is presented in a table containing from one to five levels of service, ranging from 'lowest' through 'low', 'medium', 'high', to 'highest'. In some cases the increasing levels of service will be additive or cumulative, i.e. the next highest service will say "Add: [service]" meaning that at this level of service the [service] is to be added to the service available at the level immediately below it.

The table also includes a reference to the 'engineering form' of the service description in Section 2 above. As noted above, not all of the 'engineering form' requirements / levels of service will map to user-oriented service descriptions in this section. There can also be cases where more than one user-oriented service description will be mapped to a requirement / levels of service in Section 2. These are cases where facets of a service that may be of significance from the user point of view are not separable or resolvable as separate requirements / levels of service in the reference model for cost estimation.

The levels of service are presented in the context of the general data service provider reference model. An evaluation of which services at what levels are applicable to an actual real-world data service provider would depend on an analysis of its ESE role and mission, the needs of its user community, the data it supports, etc. A general indication of how the services and levels of service will vary is provided in White Paper 6, "ESE Logical Data Service Provider Types".

Ingest Service

This section presents a users' view of ingest services provided by the data service provider. A user of a DSP's ingest service might be another provider that generates products and delivers them to the DSP to be ingested, archived, and distributed.

1. Product Ingest Service

Level	Service - What does the user see?	Reference: 2.1 a
Lowest	The DSP will occasionally accept products on an ad-hoc basis.	
Low	The DSP will accept products on an ad-hoc basis.	
Medium	The DSP will accept products on an ad-hoc basis and will return verification of receipt and successful ingest (i.e., including verification of data quality and integrity).	
High	The DSP will accept products on a routine basis and will return verification of receipt and successful ingest.	
Highest	The DSP will accept products on an operational, time-critical basis and will return immediate verification of receipt and successful ingest.	

In practice, the requirement for a particular level of service for ingest will vary from data stream to data stream; a real-world data service provider will ingest multiple data streams or flows of products at different levels of service. A distinction between data service providers will be the highest level of service they will provide for any data stream.

Processing Services

This section presents a users' view of the processing, or product generation, services provided by a data service provider. Users of a DSP A's processing service might include:

1) another DSP whose own product generation might be dependent upon products being generated on a timely basis by DSP A;

2) a flight project team who 'subcontracted' routine generation of products from its mission data to DSP A;

3) a research project team who turned to the DSP to generate research products from DSP-held data on an ad-hoc basis;

4) a researcher involved with a field experiment who depends on DSP products as aids in conducting the field campaign.

In the second and third cases, the user might wish to provide the science algorithms according to which a product is generated, perhaps retaining responsibility for science quality assurance and for judging the readiness of a product for general use.

In the discussion of processing services, two types of products are referred to, ‘standard’ products and ‘ad hoc’ products. Standard products are products that are produced over a period of time using a validated processing algorithm that gains both stability and peer acceptance, such that a science or applications user of the product can rely upon it and use it with confidence. Other products might be produced for a specific study, or might be research products of uncertain quality whose limitations a user would need to take into account.

A data service provider might offer at least three distinct modes of production to its users:

1) operational generation of new products, whether scheduled (perhaps to keep up with data inflow), or on demand, characterized by high reliability and robustness, perhaps with terms of service formally agreed to a level of service agreement or equivalent;

2) non-operational, where new products are generated without a fixed schedule or guarantee of responsiveness, perhaps using DSP resources on an as available basis;

3) reprocessing, generating new version(s) of previously generated products, either on request or according to a schedule negotiated with the user (where the user might be a flight project or science team responsible for the science algorithm software used to generate the products.

1. Operational Product Service

Level	Service - What does the user see?	Reference: 2.2 a
Lowest	The DSP will produce a product within 30 days of receiving its inputs.	
Low		
Medium	The DSP will produce a product within 7 days of receiving its inputs, by a schedule or in response to an on-demand request.	
High		
Highest	The DSP will produce a product within 2 days of receiving its inputs, by a schedule or in response to an on-demand request.	

The operational product service is generally associated with standard products. A real-world data service provider will provide a mix of levels, depending on the requirements for each of its products. A distinction between data service providers will be the highest level of service they will provide for any production stream.

2. Non-Operational Product Service

Level	Service - What does the user see?	Reference: 2.2 b
Lowest	The DSP will produce non-operational products on a time available basis - no goals or targets.	
Low		
Medium	The DSP will produce non-operational products, meeting general goals negotiated with the user.	
High		
Highest	The DSP will produce non-operational products, meeting specific targets negotiated with the user.	

3. Reprocessing Service

Level	Service - What does the user see?	Reference: 2.2 c, d
Lowest	The DSP will reprocess products per user request, on a time available basis - no goals or targets.	
Low		
Medium	The DSP will reprocess products, meeting general goals negotiated with the user.	
High		
Highest	The DSP will reprocess products, according to a reprocessing schedule negotiated with the user.	

The reprocessing service is generally associated with standard products, where further research causes a processing algorithm to be improved to the point where scientists recommend reprocessing to produce a new version of the product using the improved algorithm. Reprocessing of a product might also be driven by reprocessing of one or more of the inputs used to produce it, or of corrections or improvements to ancillary data.

4. Science Software Integration and Test Service

Level	Service - What does the user see?	Reference: 2.2 e
Lowest	The DSP will not accept user provided software.	
Low		
Medium	The DSP will accept science software from a user for standard products to be produced by the DSP, and perform integration test and verify readiness of the software for production.	

High	Add: The DSP will accept science software from a user for research products to be produced by the DSP, and perform integration test and verify readiness of the software for production.
Highest	Add: The DSP will accept science software from a user for data integration or data mining tasks, and perform integration test and verify readiness of the software for production.

Documentation Service

This section presents a users' view of the product documentation services provided by a data service provider. All users of a DSP's products will rely on the DSP's documentation to support their understanding and use of the products. Users such as a flight project or research project that provides products to a DSP for archive and distribution, or that 'subcontracts' product generation to a DSP, will be concerned with the quality of DSP documentation, especially if members of their team will have to use it. The eventual long-term archive for the DSP's data and products will also have a vital concern with the quality of DSP documentation.

Documentation includes complete documentation of all product generation software used by the DSP.

1. Documentation Service

Level	Service - What does the user see?	Reference: 2.3 a, b, c
Lowest	The DSP provides GCMD DIFs describing its data and products, and other documentation either only as received from data or product sources or informal documentation of products it generates.	
Low	Add: The DSP provides read software and format documentation for its data and products.	
Medium	Add: The DSP provided users' guides for its data and products (product type level, product instance (a.k.a. granule) level, electronic or hard copy, including journal references).	
High	Add: The DSP includes user feedback about data and products in its documentation.	
Highest	Add: The DSP ensures that its data and products are documented to the standard required for long term archiving.	

Archive Services

This section presents a users' view of the archive service provided by a data service provider. For many DSP users, especially researchers or applications specialists who obtain data or products from the DSP, the archive service will be out of sight, a layer below the access services they interact with directly. They will rely implicitly on the archive service as they do on the DSP's sustaining engineering and other vital but not directly visible services. Other users of the

DSP's archive service will include flight projects or research teams who will rely on the DSP to archive their data and products, and while they would accomplish delivery of data and products to the DSP via the DSP's ingest service, they will be concerned with the level of archive service the DSP will provide for their data and products. What would be visible to the user in this sense is the commitment of the DSP to provide a given level of service (perhaps captured in a formal agreement).

1. Archive Quality Monitoring Service

Level	Service - What does the user see?	Reference: 2.4 e, f
Lowest	Quality screening (read after write) on data entering the archive;	
Low	Add: 1% per year random screening to detect and replace degrading media;	
Medium	Add: 5% per year random screening;	
High	Add: Exit screening (tracking of archive read failures, corrected errors, etc., to find and replace degrading media);	
Highest	Add: 10% per year random screening;	

A user who would see reports of the quality monitoring might be a flight project or research group that had entrusted the data service provider with the stewardship of data critical to its work. A 'regular' end user would not see the effects of this service directly. The next service, archive backup service, is a similar case.

2. Archive Backup Service

Level	Service - What does the user see?	Reference: 2.4 h, i
Lowest	Partial on-site backup, with regular sampling to verify integrity of the backup.	
Low	Add: Media migration on ad hoc basis when needed.	
Medium	Add: Partial off-site backup, with regular sampling to verify integrity of the backup.	
High	Add: Full off-site backup, with regular sampling to verify integrity of the backup.	
Highest	Add: Media, archive technology migration planned and budgeted for.	

Search and Order Services

This section presents a users' view of the search and order service provided by a data service provider.

1. Availability of Search and Order Service

Level	Service - What does the user see?	Reference: 2.5 a
Lowest	Search and Order service is open to a limited team of scientists or applications specialists *with at least a minimal capability for access to the science and applications community and the public*.	
Low		
Medium	Search and Order service is open to the science and/or an applications community *with at least a minimal capability for public access*.	
High		
Highest	Search and Order service is public, open to all users.	

Public access may be limited in some cases by constraints levied by the original data source. Data service providers with a highly focused primary mission (such as a flight project instrument team) may meet the requirement for public access by teaming with other data service providers actively engaged in public access and services.

2. Search Service

Level	Service - What does the user see?	Reference: 2.5 b
Lowest	Search a list of available DSP products for ones of interest.	
Low	Search for DSP products of a specified type by time and space coverage.	
Medium	Search for any/all DSP products by time and space coverage.	
High	Add: search for any/all DSP products by geophysical parameter, time, and space *(by named spatial object from a catalog as well as by coordinates)* coverage.	
Highest	Add: search for any/all DSP products pertaining to a given phenomenon (e.g. hurricane, volcanic eruption, El Nino), campaign, research project.	

3. Search Service - Product Descriptions

Level	Service - What does the user see?	Reference: 2.5 c
Lowest	Option to view standard guide and DIF metadata for any product found by a search.	
Low	Add: option in some cases to view a few references to published papers that describe the production or use of the product.	
Medium	Add: option in most cases to view a few references to published papers that describe the production or use of the product.	

High	Add: option in some cases to view detailed algorithm and use explanations for the product.
Highest	Add: option in most cases to view detailed algorithm and use explanations for the product.

3. Search and Order Service Mode

Level	Service - What does the user see?	Reference: 2.5 b, d
Lowest	Web accessible user interface for search and order for the DSP.	
Low		
Medium	Web accessible user interface for search and order for this and other ESE DSPs.	
High		
Highest	User system accessible interface for automated search and order.	

Access and Distribution Services

This section presents a users' view of the access and distribution service provided by a data service provider.

1. Availability of DSP Data and Product Holdings

Level	Service - What does the user see?	Reference: 2.6 a
Lowest	DSP data, products, *and documentation* are available to a limited team of scientists or applications specialists, *with at least a minimal capability for access to the science and applications community and the public*.	
Low		
Medium	DSP data, products, *and documentation,* are available to the science and applications community, *with at least a minimal capability for public access*.	
High		
Highest	DSP data, products, *and documentation* are public, open to all users	

Data service providers with a highly focused primary mission (such as a flight project instrument team) may meet the requirement for at least minimal public access by teaming with other data service providers actively engaged in public access and services.

In the case of a real-world data service provider, product availability can vary on a product by product basis. Some products (such as those obtained from international or commercial sources) might have general distribution restrictions that NASA accepts in order to secure access for approved NASA scientists. Other products might go through a period of limited access while they are validated and corrected or improved before being approved for general access. What may distinguish data service providers from one another is the highest level of availability they are willing or required by their ESE role to support.

2. Access and Distribution Service Mode

Level	Service - What does the user see?	Reference: 2.6 d, e
Lowest	Data available on a provider scheduled basis, no flexibility for user.	
Low	Data available to user on a request basis.	
Medium	Add: data available to user on a subscription (i.e. standing order) basis.	
High	Add: data available to user on an operational basis (DSP will commit to terms of service for scheduled or on demand distribution).	
Highest	Add: user’s system can access data from DSP system directly by network.	

3. Data Services

Level	Service - What does the user see?	Reference: 2.6 c
Lowest	Reformatting available for a few selected DSP products.	
Low	Reformatting, subsetting, available for less than half of DSP products.	
Medium	Subsetting, reformatting, packaging available for less than half of DSP products.	
High	Subsetting, reformatting (including GIS support), packaging available for more than half of DSP products.	
Highest	Subsetting, reformatting (including GIS support), resampling, reprojection, packaging available for most DSP products.	

The specific data services available for a particular product will depend on the nature of the product and the needs of its users; the higher levels of service mean a progressively wider variety of data services available for a progressively increasing fraction of the data service provider’s holdings.

4. Access and Distribution Service Timeliness

Level	Service - What does the user see?	Reference: 2.6 f, g
Lowest	Availability of a product for network delivery (e.g. staged for FTP pick up or push) within twenty four hours of request (or of production of a product ordered in advance). Shipping of a product on one form of media within one month of request.	
Low	Add: Shipping of a product on media, with user's choice from several media types, within one week of request.	
Medium	Add: Availability of a product for network delivery within ten minutes.	
High	Add: Availability of a product for network delivery within ten seconds (e.g. data or products held on-line). Shipping of a product on media within three days of request.	
Highest	Add: Availability of a product for automated, direct access via network within ten seconds (e.g. data or products held on-line for access by user software).	

In the case of a real-world data service provider, the level of service will almost always vary with the product, depending in each case on factors such as user demand and sheer size. Also, these service levels are defined in terms of a request for a single product. A request for a year's worth of a product, especially a large low level product, e.g. MODIS level 1, would take much longer, and would be negotiated with the user.

User Support Services

This section presents a users' view of the user support service provided by a data service provider.

1. Availability of User Support

Level	Service - What does the user see?	Reference: 2.7 c, d
Lowest	Casual telephone or email contact.	
Low	Add: Help Desk staffed during local work day (5 days x 8 hours per day).	
Medium	Add: On-line help, FAQ, data / product and service descriptions.	
High	Add: Help Desk staffed during work day for all U.S. (5 days x 12 hours/day).	

Highest	Add: Help Desk staffed during work day worldwide (5 or 7 days x 24 hours per day).

2. Capability of User Support Staff

Level	Service - What does the user see?	Reference: 2.7 b
Lowest	Basic knowledge of what data and products are available, what network / media / delivery options exist.	
Low	Add some knowledge of format detail for most popular products.	
Medium	Add comprehensive knowledge of format details for most if not all products.	
High	Add technical expertise in data structures, use of tools for format conversion, subsetting, analysis, etc..	
Highest	Add science expertise in data / product quality and research uses for data and products.	

3. Outreach to Potential New Users

Level	Service - What does the user see?	Reference: 2.7 e
Lowest	None.	
Low	DSP outreach material (pamphlets, brochures, posters, etc.)	
Medium	DSP's booth or booth participation at conferences, at least four times/year.	
High	Add: DSP staff providing or contributing to workshops and/or user training sessions at conferences.	
Highest	Add: DSP staff providing user training sessions at universities, schools, etc.	

Applications Software Service

This section presents a users' view of the application software service (a subset of implementation as defined above) provided by a data service provider. Users, except possibly for an advisory panel, will not see or interact with the initial implementation of the data service provider's system capabilities or their technology refresh or expansion. Users will see applications software developed by the data service provider to meet users' needs.

1. Applications Software Service

Level	Service - What does the user see?	Reference: 2.12 e
Lowest	None.	
Low	DSP developed software tools to read DSP products.	
Medium	DSP developed software tools for use by users to unpack, subset, or otherwise manipulate DSP products.	
High	Add: DSP developed product generation software embodying science algorithms, e.g. to produce a product to meet a particular user need.	
Highest	Add: DSP developed applications software to perform a 'data mining' or data integration operation to meet a user need.	

Appendix A - Draft Program Level Requirements

This section contains the set of program level requirements drafted by the SEEDS Formulation Team in September, 2001, as "NewDISS Level 0 Requirements", with the new term "SEEDS" replacing "NewDISS". The cost model requirements template that follows fits within the general framework of the program level requirements in this section.

A.1 General Requirements

a. Data service providers will fully participate (TBD) in SEEDS community-based management processes including standards and interface determination, reuse/architecture refinement, metrics collection, and Enterprise peer review.

b. All data service providers will comply with SEEDS Level of Service requirements for core functions and data products (TBD) and will adhere to SEEDS required core interfaces and standards (TBD). Deviation from core standards must be requested and approved via the SEEDS waiver process (TBD).

c. Data service providers will provide metrics (TBD) on data production and utilization to the SEEDS Office on a routine (TBD) basis.

d. Data service providers and projects will participate in an annual (TBD) broad-based peer review of ESE data management activities.

e. ESE Mission Projects will produce a Life Cycle Data Management (LCDM) Plan. Changes to the LCDM plan will be approved by the SEEDS Office (TBD).

f. To the extent possible and where cost effective, data service providers will reuse software and system components developed by previously NASA funded activities. Projects will enable possible reuse of their software available by following the system design guidelines provide by the SEEDS reference architecture (TBD).

A.2 General Science Requirements

a. Data service providers will provide support to and receive technical direction from an appropriate NASA ESE science parameter team.

b. Principal Investigators will propose a suite of standard science products subject to peer review approval of an Algorithm Theoretical Basis Document.

c. Each data service provider will have a Science Advisory Group that will review progress and plans on a routine basis.

A.3 Production, Archive, and Distribution Requirements

a. All raw data will be acquired will be calibrated and geolocated to a reference sphere. Calibrated and georeferenced data will made available to all users.

b. Data at the "raw" sensor level (NASA Level 0 plus appended calibration and geolocation information) must be archived permanently.

c. All standard science data (Level 1b, Level 2, and Level 3) produced will be made available to any user who requests it without discrimination.

d. All standard data products available to a science team member will be made available to general science users.

e. All standard data produced will be archived until the end of the science mission or until transfer to an approved permanent archive.

f. Data service providers will receive orders for data products from the general public and will fulfill those orders with an average delivery time (elapsed time between when the order was completed and product was shipped) of less than five working days.

A.4 Standards and Interface Compliance

a. Metadata for all standard products will be produced in accordance with the SEEDS core metadata standard.

b. Metadata for all archived standard data products must be searchable by spatial and temporal extent, and must be locatable by the general user via the world wide web.

c. Standard data products made available to the LTA, to another SEEDS data service provider and to users will be available in one of the SEEDS core formats.

d. All standard data products will be cataloged in the Global Change Master Directory (GCMD). Data service providers will provide Directory Interchange Format (DIF) documents on all standard data products to the GCMD prior to release of the data products.

References and Acronym List

The References Section and the Acronym List for all of these Working Papers is in the document

“References and Acronyms for the Levels of Service / Cost Estimation Working Papers ”.

SEEDS

Working Paper Six:

ESE Logical Data Service Provider Types

April 24, 2002

G. Hunolt, SGT, Inc.

Outline

1.0 Introduction

2.0 Logical Data Service Provider Types in Relation to the Data Service Provider Reference
Model
2.1 Logical Data Service Provider Types as Reference Model Subsets
2.2 Using Logical Data Services Provider Types

3.0 ESE Logical Data Service Provider Types
3.1 Backbone Data Center
3.1.1 Backbone Data Center Concept
3.1.2 Backbone Data Center Functions
3.2 Mission Data Center
3.2.1 Mission Data Center Concept
3.2.2 Mission Data Center Functions
3.3 Science Data Service Provider
3.3.1 Science Data Center Concept
3.3.2 Science Data Center Functions
3.4 Systematic Measurements Center
3.4.1 Systematic Measurements Center Concept
3.4.2 Systematic Measurements Center Functions
3.5 Applications Center
3.5.1 Applications Center Concept
3.5.2 Applications Center Functions
3.6 Information Center
3.6.1 Information Center Concept
3.6.2 Information Center Functions
3.7 Long Term Archive Center
3.7.1 Long Term Archive Center Concept
3.7.2 Long Term Archive Center Functions

4.0 Allocation of Requirements / LOS to Logical Data Service Provider Types
4.1 Ingest
4.2 Processing
4.3 Documentation
4.4 Archive
4.5 Search and Order
4.6 Access and Distribution
4.7 User Support
4.8 Instrument / Mission Operations
4.9 Sustaining Engineering

4.10 Engineering Support
4.11 Technical Coordination
4.12 Implementation
4.13 Management
4.14 Facility / Infrastructure

References and Acronym List

Introduction

This working paper is the sixth of a set of papers that describes the SEEDS (Strategic Evolution of Earth Science Enterprise Data Systems) Levels of Service (LOS) / Cost Estimation (LOS/CE) study. The study goal is to develop a cost estimation model and coupled requirements and levels of services to support the SEEDS Formulation team in estimating the life cycle costs of future ESE data service providers and supporting systems, where 'data service provider' is used as a generic term for any data/information related activity. The set of working papers is intended to serve as a vehicle for coordinating work on the project, obtaining feedback and guidance from ESDIS SOO and the user community, and as embryos of reports that will be produced as the task proceeds.

As working papers, each version of each paper that appears represents a snapshot in time, with the work in various stages of completion. As work progresses the content (and sometimes the organization) of the working papers will change reflecting progress made, responses to feedback and guidance received, etc.

This sixth working paper of the set describes the current set of logical data service provider types developed for the LOS/CE study, and reflects results of the February, 2002, SEEDS Community Workshop.

Section 2 discusses the logical data service providers as subsets of the general data services provider model.

Section 3 presents descriptions of the logical data service provider types.

Section 4 presents a mapping of the reference model requirements and levels of service from Working Paper 5, "Data Service Provider Reference Model - Requirements / Levels of Service" to the logical data service provider types in Section 3.

Logical Data Service Provider Types in Relation to the Data Service Provider Reference Model

The Data Service Provider Reference Model is a general functional model of an abstract, generic data service provider that includes a full set of functional areas. The model is described by the set of functional areas, corresponding requirements and levels of service, and a parameter set that is mapped to the functional areas and levels of service. (See Working Paper 3, "Data Service Provider Reference Model - Functional Areas", Working Paper 4, "Data Service Provider Reference Model - Model Parameters", and Working Paper 5, "Data Service Provider Reference Model - Requirements and Levels of Service".)

Logical Data Service Provider Types as Reference Model Subsets

The general data service provider reference model includes all functions / areas of cost that a generic data service provider might perform. While an actual working data service provider could conceivably perform all of the functions included in the model, most if not all actual data service providers perform a subset of them, e.g. most providers will not have a requirement in the area of instrument / mission operations. Many well known actual data centers such as the NASA Distributed Active Archive Centers or the NOAA national data centers perform a subset of the general set of functions. Some data service providers, e.g. MODAPS (the MODIS Adaptive Processing System, a sample of a science team processing facility that does not perform archive or general user distribution), are different in function from many well known data centers but fit within the framework of the data service provider reference model.

The Cost Estimation Tool will allow a planner, for example one planning a data service to support a flight project, to:

1) select those functions that are required for his/her particular mission (in effect creating a 'custom' subset of the model);

2) specify the particular mission requirements the real instantiation of it must meet (e.g. data volumes to be ingested, processed, stored, and/or distributed);

3) produce an estimated cost for implementing and operating it.

To facilitate overall ESE data service architecture studies (where a 'data service architecture' is a collection of data service providers and the interconnections between them), a set of 'logical data service provider types' has been defined. Each of these types is a functional subset of the general reference model organized around a defined class of ESE role or mission. These are 'logical' types in that there is no explicit or implicit 1:1 mapping of an instance of a logical data service provider type to a physical entity. While some actual data service providers might match a logical type, most will perform the functions of more than one logical type, and may also perform multiple data service activities within the scope of a type (such as a DAAC that performs archive and distribution for several flight projects). Because the logical data service provider types are only a few of the possible subsets of the general model, they constitute an open set to which additions (and subtractions) can be readily made as needed to facilitate architecture trade studies or other uses.

Using Logical Data Service Provider Types

The logical data service provider types can be used in two ways, as discussed in Working Paper 2, "Cost Estimation by Analogy Model", which includes scenarios showing how the Cost Estimation Tool would be used. Two different modes of use of the tool are described.

The first mode is to produce a life-cycle cost estimate for a particular data service provider activity to be performed by a new provider or as an additional task by an existing provider. In this case the user of the tool would select the data service provider functions needed in the particular case, and produce an estimate of the cost for implementing and operating it. For this purpose, while a user would have the freedom to create a custom set of functions (in effect creating a custom subset of the general data service provider model) the user would also have the option of deciding that his/her needs corresponded to a logical data service provider type and using a template for it to facilitate producing the cost estimate.

The second mode is to produce an overall estimate for an ESE architecture, some combination of organizations performing data services functions such that the aggregate ESE requirements for data services are met. As described in Working Paper 2, producing a cost estimate for an overall ESE architecture, or a number of different estimates for alternative architectures, requires the user to deal with a large number of data service activities, having to encompass the ESE as a whole. The user in this situation will not be able to consider any one activity exhaustively, and for this purpose the logical data service provider types will be of great help.

In either case, the logical types must be useful subsets of the general reference model, i.e. they must be organized around an ESE role or mission that is significant in the real world. Only then will they be of genuine value to the ESE data services architect, or attractive to the individual or team planning a single activity.

ESE Logical Data Service Provider Types

This section describes the current set of ESE logical data service provider types, drawing on the NewDISS concept paper “Draft Version 1.0 - NewDISS: A 6-to-10-year Approach to Data Systems and Services for NASA’s Earth Science Enterprise”, October 2000 for a starting point. For each logical data service provider type, this section will present the conceptual description taken from the concept paper and a description of the functions of the data service provider type in terms of the data service provider reference model and its functional areas (defining the subset of the reference model that applies to the data service provider type).

The NewDISS concept paper introduces its discussion of NewDISS data service provider types: “NASA’s ESE has requirements for collection and synthesis of scientific information, for bringing synthesized data products to bear on unanswered scientific questions, and for preserving data and information for future scientific discovery. … NewDISS is therefore seen as consisting of a dynamic network of interconnected components, each responsive to its environment, containing capabilities for change over time through feedback with the science community. These components will be responsible for executing NewDISS data management functions and must allow easy participation by scientists and data and services providers. The components of NewDISS have been conceptualized (October, 2000) as including “Backbone” processing centers, PI-managed Mission Data Centers, Science Data Centers, and Multi-Mission Centers [here Systematic Measurement Centers].”

Three additional data service provider types are added:

1. Applications Center, focused on uses and users other than research, given the existence of NASA funded applications activities such as Type III Earth Science Information Partners (ESIPs) and Regional Earth Science Applications Centers (RESACs);

2. Information Center, focused on information describing data and products rather than the data and products themselves, based on discussion at the Formulation Team Retreat, November 7-8, 2001, where ‘Echo’ was suggested as a possible future instance, and the Global Change Master Directory (GCMD) is plainly a currently operational instance;

3. Long Term Archive Center, focused on permanent preservation and archiving of data and products and their documentation and active support to climate research, etc., based on a request from Matt Schwaller, a member of the Formulation Team and leader of the Earth Science Data Life-Cycle study. Long term archiving is strictly speaking not an ESE responsibility, but inclusion of a hypothetical Long Term Archive data service provider type is intended to support planning that NASA is doing with NOAA and USGS, the agencies who have (with the National Archives and Records Administration, NARA) the long term archive responsibility.

A “data service provider” does not necessarily imply a physically distinct institution. An institution such as a NASA center, a university, an organization of another US Government Agency such as USGS or NOAA can host a data service provider or a combination of data service providers. This is equivalent to the existing situation in which the University of Colorado hosts the National Snow and Ice Data Center (NSIDC) DAAC, or the USGS’s EROS Data Center (EDC) hosts the EDC DAAC.

Backbone Data Center

This section describes the generic Backbone Data Center type.

Backbone Data Center Concept

The following is the concept for Backbone Data Centers, from the NewDISS Concept Paper: “These centers, most likely evolving from some of the current DAAC’s, will address NASA’s responsibility for preserving and protecting the large volumes of data from the ESE satellite missions. One of the primary roles of the backbone data centers will be to preserve the basic data. Clearly, NASA can provide a considerable amount of existing infrastructure and technical skill needed to provide satellite mission data downlink and “level 0” or “level 1” data processing. Teaming NASA missions with Backbone Data Centers in the Announcement of Opportunity (AO) process for backup or for generation of basic data products may well be an attractive option for handling some of the core data management requirements of NewDISS. Another role for the Backbone Data Centers will be to acquire products agreed to be scientifically important for preservation and to prepare all these data for long-term archiving. These data centers will need to address network connectivity as part of their on-going activities. Selection of these services will be driven by PI-teaming arrangements, using either NASA-available resources or competitive alternatives. Backbone Data Centers, staffed by professional data managers, provide a core set of historical experience and proven capabilities. As such, they provide a means for risk mitigation against the failure of one or more of the NewDISS components by serving as backup centers for the other parts of the NewDISS. These data centers would most likely be few in number to ensure the cost-effectiveness of the NewDISS.”

Backbone Data Center Functions

In general, the Backbone Data Center is expected to provide stable and highly robust services, with a key responsibility for data preservation and documentation, and with a mandate to provide professional data management as a resource for ESE as a whole. A Backbone Data Center is not identified with a particular mission or project but provides data management services in support of multiple missions and the NASA science program in general. The Backbone Data Center would have an indefinite lifespan subject to regular independent review of its performance.

In addition to its operational functions, the Backbone Data Center would bring a high level of expertise in the areas of data stewardship, science, and information technology in planning and conducting its own activities, working with its science user (and other as appropriate) community to understand how best to meet its needs, and coordinating with ESE and other ESE data service providers in many areas (standards, interfaces, interoperability, data interuse across providers, best practices, data stewardship, user support, information technology, etc.).

The Backbone Data Center would be expected to move to increasingly automated services, e.g. allowing user software to interact directly with center's system to provide the functional equivalent of a manual search and order system and direct access to data and products. Backbone Data Centers may also increasingly support data integration and data mining for users, using software developed by the center for the purpose or in some cases using software provided by the user.

The Backbone Data Center could provide processing functions for NASA missions through teaming arrangements with NASA Principal Investigators, and can serve as a backup to other ESE data service providers.

The paragraphs below will briefly discuss the Backbone Data Center role in each of the general data service provider reference model's functional areas.

Ingest - The Backbone Data Center performs ingest of a wide variety of data types, ranging from low level data streams to ancillary data to all of the levels of derived products, including their metadata, documentation, etc. In some cases the ingest function must be performed on a time critical, operational basis, e.g. for data and supporting information received from operating satellite platforms via NASA or other agency mission operations and communications systems. Level of service agreements or equivalent (e.g. operations agreements, interface agreements) with sources may be required. Quality control on incoming data is especially critical for lower level (e.g. level 0) data ingested, as the Backbone Data Center must detect bad data and request replacement data from operational sources that may have a limited capability for storing and retransmitting data.

Processing - The Backbone Data Center may perform processing through a teaming arrangement with a flight mission Principal Investigator, which can include large scale (in terms of number of products generated and /or product volume data) operational, schedule driven 'standard product' processing and reprocessing, perhaps with emphasis on Level 1 processing vs higher level derived product processing. The Backbone Data Center would provide a science software integration and test service. Operational processing by the Backbone Data Center would be highly reliable with tight quality control.

The Backbone Data Center may also perform non-operational processing which could include generation of research products, data integration products, and data mining.

Documentation - The Backbone Data Center ensures that its data and product holdings are documented to the SEEDS adopted standard for long term archiving, working as necessary with external data sources (e.g. other data service providers) to capture all needed information. The Backbone Data Center also ensures produces search metadata, product guides, etc., to SEEDS adopted standards, which would likely include Federal Geographic Data Committee (FGDC) compliant metadata.

Archive - The Backbone Data Center provides a very robust archive capability, performing insertion of data into archive storage, and performing archive quality data stewardship, including preservation of data, metadata, and documentation within the archive. Preservation measures should include quality screening of data entering and exiting the archive, quality screening of archive media, off-site backup with sampling and tested restoration to verify integrity, and accomplishing migrations from one type of media to another.

Search and Order - The Backbone Data Center serves a broad user community with a robust and flexible search and order capability that supports user interaction with search and order services and, increasingly, supports automated search and order interaction between software running on a user system and the Backbone Data Center system. The search capability allows a user to apply criteria that might include geophysical parameter(s), spatial-temporal coverage, specific product names, etc., to the metadata describing available data and products and returning to the user listings supplemented by descriptive information of those data or product types and instances that meet the criteria. The 'order' capability includes a request/permission step, regardless of how implemented (e.g. manual or automated), where a request for a set of data or product instances, perhaps the results of (or a selected subset of the results of) a search, is processed and accepted or denied.

Backbone Data Center search and order can include providing local user interface and capability and/or providing an interface to a broader based, ESE cross-site search and order capability.

Access and Distribution - The Backbone Data Center serves a broad user community with a robust access and distribution (electronic and media) service, including offering data services such as subsetting, reformatting, reprojecting, packaging in response to user needs. The Backbone Data Center will increasingly support automated access to its data and products by user software.

The Backbone Data Center will also transfer data and documentation to designated long term archive centers in accordance with life cycle data management plans.

User Support - The Backbone Data Center provides effective user support for a wide range of users, including support provided in direct contact with users by user support staff, e.g. responding to queries, taking of orders, staffing a help desk (i.e., staff awaiting user contacts who can assist in ordering, track and status pending requests, resolve problems, etc.), etc. The Backbone Data Center will increasingly offer more automated user support aids (beginning with on-line documentation, FAQ, etc.) to meet increasing demands on user support with the proliferation of data types, data sources, and tools for users. User support staff should include science expertise to provide users with assistance in selecting and using data.

The Backbone Data Center coordinates its user support with other ESE data service providers (e.g. for user referral services). It performs outreach to potential new users, and participates in coordinated outreach activities with other ESE data service providers.

Instrument / Mission Operations - The Backbone Data Center does not perform this function.

Sustaining Engineering - The Backbone Data Center performs sustaining engineering, with no or very infrequent interruption of operational capabilities.

Engineering Support - The Backbone Data Center performs engineering support functions with no or very infrequent interruption of its operations.

Technical Coordination - The Backbone Data Center is heavily involved in technical coordination. It participates in SEEDS system level processes, including coordination on data management, documentation standards, data stewardship (including standards for content of life cycle data management plans), standards and best practices (including quality assurance standards and practices), interfaces, common metrics, and interoperability (e.g. for data access and integration), across / within SEEDS and with other systems and networks as needed to support the ESE program.

The Backbone Data Center participates in ongoing examination of the changing needs of the ESE science and applications program and the consequent impacts on the roles, missions, and services of ESE data service providers.

The Backbone Data Center participates in coordination of user support guidelines and practices across the network of ESE data service providers and with other data centers as needed to support the ESE science and applications program.

The Backbone Data Center cooperates with other ESE data service providers in representing ESE / SEEDS in broader community processes in areas such as standards, interoperability, data management, security, etc.

The Backbone Data Center also participates in SEEDS level and/or bilateral processes to coordinate production and delivery of products between itself and other ESE data service providers.

Implementation - The Backbone Data Center develops the data and information system capabilities it requires to perform its mission, including initial design and implementation of the data system (hardware and system software) and applications software and expansion or replacement (i.e. technology refresh) as needed over its operating life.

The Backbone Data Center also maintains an ongoing applications software development effort. Applications software can include software to perform data services (e.g. subsetting, reformatting, reprojection, etc.) for more of its products, software tools for use by users to unpack, subset, or otherwise manipulate products provided by the Backbone Data Center, product generation software embodying science algorithms, e.g. to produce a product to meet a particular user need, and to perform a ‘data mining’ or data integration operation to meet a user need.

Management - The Backbone Data Center performs a variety of site-level management functions as well as performing direct management of its functional areas.

Site-level management by the Backbone Data Center includes planning information technology upgrades / technology refreshes, based on assessments of changing mission or user needs and availability of new technology. It includes developing data stewardship practices, performing data administration with science advice (via the User Advisory Group and other appropriate bodies), developing and maintaining life cycle data management plans (which address data migrations). It also includes coordinating its internal science activities and its interaction with the ESE and broader science community, including a visiting scientist program or equivalent, collaboration among ESE data service providers to support science needs, annual Enterprise peer review, and support for its User Advisory Group (which includes representation from the science, applications, education, etc., communities that it serves) and any other ESE or broader advisory activities that may be appropriate.

The Backbone Data Center also participates in ESE / SEEDS management processes, strategic planning, and coordination with other data centers and activities beyond ESE/SEEDS.

Facility / Infrastructure - The Backbone Data Center provides and maintains a fully furnished and equipped, environmentally controlled, physically secure facility to house its staff, systems, and data and information holdings, including a separate off-site backup facility for its data and information holdings. The Backbone Data Center ensures system and site security according to established NASA security policies and practices.

The Backbone Data Center performs resource planning, logistics, supplies inventory and acquisition, and facility management. It provides for purchase of supplies, facility lease and utility costs and other similar overhead costs, hardware maintenance, COTS licenses, etc.

The Backbone Data Center provides facility / infrastructure support at a level that ensures no or very infrequent interruption of its operations.

Mission Data Center

This section describes the generic Mission Data Center type.

Mission Data Center Concept

The following is the concept for NewDISS Mission Data Centers, from the NewDISS Concept Paper: “These data systems are specifically affiliated with instruments or satellite systems. They are either PI-led or facility/project-led. They provide key measurements and standard products from NASA-supported satellite instruments. The key characteristic of the mission data centers is that they will be engineered and implemented as part of an ESE mission proposal. It is anticipated that these Mission Data Centers could leverage the activity at the current ESE data management infrastructure: the ECS flight operations and science data systems and the other hardware and software infrastructure at the DAAC's, the ESIP's, and the SCF's. These data centers will need to address network connectivity as part of their on-going activities. Selection of these services will be driven by PI-teaming arrangements, using either NASA-available resources or competitive alternatives. Mission Data Centers will also need to address satellite/instrument command and control and data downlink. Selection of these services will be driven by PI-teaming arrangements, using either NASA-available resources or competitive alternatives, such as commercially provided or university support services."

Mission Data Centers will be responsible for their data management functions during an Earth-observation space flight mission. These data service providers will be funded by the mission selected through the ESE flight programs and will be selected by competitive selection for future ESE missions."

Mission Data Center Functions

In general, the Mission Data Center is an element of a particular ESE mission that exists to provide data management services for the life of that mission. The mission might be might involve an instrument on an independently operated spacecraft (such as SeaWinds on a Japanese platform) or might include multiple instruments on a dedicated spacecraft (such as Terra or Aqua). The services provided by the Mission Data Center extend from instrument or platform command and control through generation and distribution to mission science team members of science products derived from instrument data for quality assurance, validation, and research. A Mission Data Center would provide instrument data and science products to a Backbone Data Center for distribution to the broad user community and archive after the mission life is completed.

Ingest - The Mission Data Center ingests instrument and spacecraft telemetry and instrument data from NASA or other spacecraft operations and communications systems, and ancillary data needed to support product generation from various sources. Ingest of instrument data and instrument and spacecraft telemetry might be performed on a time critical, operational basis, and the Mission Data Center must detect bad data and request replacement data from operational sources that may have a limited capability for storing and retransmitting data.

Processing - The Mission Data Centers will perform small to large scale (in terms of number of products generated and /or product volume data) 'standard product' processing and reprocessing. If the processing is performed to meet the needs of the mission science team only, it can be performed as the team requires. If the processing also must meet the needs of other missions (e.g. as ancillary products), science teams, or other users, it may be performed on an operational basis (especially once processing algorithms become stable). Processing by the Mission Data Center would include tight quality control. The Mission Data Center could team with a Backbone Data

Center for the processing service, especially if there is a requirement for routine, operational generation of ‘standard’ products.

Documentation - The Mission Data Center generates complete documentation of its instrument data and all derived products. The Mission Data Center cooperates with a Backbone Data Center that receives its data after completion of its mission to ensure that documentation is brought to long term archiving standards.

Archive - The Mission Data Center would not perform an archive function per se, but would maintain secure working storage of data and products until their transfer to a Backbone Data Center at some time during the mission or after completion of the mission. The Mission Data Center would maintain an off-site back up of all data for which it is responsible, and might use the services of a Backbone Data Center for this purpose.

Search and Order - The Mission Data Center serves its mission science team with a robust and flexible search and order capability tailored to meet the needs of the science team.

Access and Distribution - The Mission Data Center provides products to the mission science team for quality assurance, validation, or research, with a search and order capability as needed to meet the needs of the mission science team. The Mission Data Center will also transfer data, products, and documentation to a Backbone Data Center either during its mission as backup or when broader distribution of its data and products is appropriate, or at the conclusion of the mission.

User Support - The Mission Data Center provides close support to member of the mission science team.

Instrument / Mission Operations - The Mission Data Center performs this function for instruments and spacecraft that are part of its mission through NASA or other appropriate operational mission management services. This includes monitoring instrument and spacecraft performance, generating instrument and (if applicable) spacecraft commands, and event scheduling.

Sustaining Engineering - The Mission Data Center performs sustaining engineering, with no or very infrequent interruption of any critical operational capabilities.

Engineering Support - The Mission Data Center performs engineering support functions as needed, but with no or very infrequent interruption of any critical operational capability.

Technical Coordination - The Mission Data Center participates in certain SEEDS system level processes, including coordination on data management, documentation standards and best practices (including quality assurance standards and practices), interfaces, security, and common metrics.

The Mission Data Center also participates in SEEDS level and/or bilateral processes to coordinate production and delivery of products between itself and other ESE data service providers.

Implementation - The Mission Data Center develops the data and information system capabilities it requires to perform its mission, including initial design and implementation of the data system (hardware and system software) and applications software and expansion or replacement as needed over its mission life.

The Mission Data Center also maintains an ongoing applications software development effort. Applications software would include ‘science software’ - product generation software embodying science algorithms, e.g. to produce a suite of products to meet the needs of the mission’s research program and the overall ESE research program. In some cases the science software would be developed to run in the Mission Data Center’s own environment, in other cases the Mission Data Center could provide science software to a Backbone Data Center for operational production.

Management - The Mission Data Center performs a variety of site-level management functions as well as performing direct management of its functional areas.

Site-level management by the Mission Data Center includes developing and maintaining life cycle data management plans for data generated by its mission, coordinating with other data service providers as needed, e.g. a Backbone Data Center to which the mission data and products and complete documentation would be transferred to after the end of the mission.

It also includes coordinating its interaction with the ESE and broader science community, collaboration among ESE data service providers to support science needs, and any ESE or broader advisory activities that may be appropriate.

The Mission Data Center also participates in ESE / SEEDS management processes, strategic planning, and coordination with other data centers and activities beyond ESE/SEEDS as needed for the success of its mission.

Facility / Infrastructure - The Mission Data Center provides and maintains a fully furnished and equipped, environmentally controlled, physically secure facility to house its staff, systems, and working storage for its data and information holdings, including a separate off-site backup facility, for which the Mission Data Center might use the services of a Backbone Data Center. The Mission Data Center ensures system and site security according to established NASA security policies and practices.

The Mission Data Center performs resource planning, logistics, supplies inventory and acquisition, and facility management. It provides for purchase of supplies, facility lease and utility costs and other similar overhead costs, hardware maintenance, COTS licenses, etc.

The Mission Data Center provides facility / infrastructure support at a level that ensures no or very infrequent interruption of its operations.

Science Data Center

This section describes the generic Science Data Center type.

Science Data Center Concept

The following is the concept for NewDISS Science Data Center, from the NewDISS Concept Paper: "These data centers will collect data from multiple missions for a user community focused on a single research question. There are several examples of these types of Science Data Centers in NASA's Space Science Enterprise. These centers are targeted at specific science questions (perhaps from the NRC Pathways Report) and/or science disciplines, and they directly support research and data analysis for specific research questions. These data centers will address network connectivity as part of their on-going activities. Selection of these services will be driven by PI-teaming arrangements, using either NASA-available resources or competitive alternatives."

Science Data Center Functions

In general, the Science Data Center is a temporary data management capability implemented to support a particular research effort by a limited community of users (which will be called its 'research team'). A Science Data Center could support a 'data mission' - supporting a research team doing science with existing data without a new flight project. The research effort could be interdisciplinary or focused on one of the traditional Earth science disciplines. The Science Data Center operates in a research environment, without the same need for robustness and performance as would be the case for an operational environment.

Ingest - The Science Data Center obtains data and products required to meet the research objectives of its research team from a variety of sources, including other ESE data service providers, other agency data centers, etc. The ingest would not be performed on a time critical, operational basis.

Processing - The Science Data Center would perform non-operational processing, and in some cases reprocessing, of new science products developed by the research team.

Documentation - The Science Data Center generates complete documentation any new science products developed by the research team that constitute new research quality products to be made available to the general science community (e.g. products cited in publications by members of the research team which should be available other scientists seeking to corroborate or extend the research performed by the team). The Science Data Center cooperates with a Backbone Data Center that receives its products after completion of its working life (or with the designated long term archive for its products) to ensure that documentation is brought to SEEDS adopted long term archiving standards.

Archive - The Science Data Center would not perform an archive function per se, but would maintain working storage of products obtained from other sources or science products generated as part of the research effort it supports.

Search and Order - The Science Data Center provides a search and order capability tailored to meet the needs of the research team it supports, perhaps supplemented to an additional capability for allowing other interested scientists to search for and order certain products the research team deems to be ready for use beyond the immediate work of the science team prior to their availability from a Backbone Data Center.

Access and Distribution - The Science Data Center will make the products collected to support the research effort readily available to members of the research team, and will perform reformatting, subsetting, or packaging of those products as needed to facilitate their interuse by the research team. The Science Data Center will also transfer new research quality science products and documentation to a Backbone Data Center when broader distribution of those products is appropriate, or at the conclusion of the research effort.

User Support - The Science Data Center provides close support to members of the research team it supports, including a help desk supplemented by on-line aids (e.g. FAQs).

Instrument / Mission Operations - None.

Sustaining Engineering - The Science Data Center performs software maintenance as needed.

Engineering Support - The Science Data Center performs engineering support functions as needed.

Technical Coordination - The Science Data Center participates in SEEDS system level processes, including coordination on data management, documentation standards, standards for content of life cycle data management plans, standards and best practices (including quality assurance standards and practices), interfaces, security, and common metrics.

The Science Data Center also participates in ESE level and/or bilateral processes to coordinate production and delivery of products between itself and other ESE data service providers.

Implementation - The Science Data Center develops the data and information system capabilities it requires by the to perform its mission, including initial design and implementation of the data system (hardware and system software) and applications software and expansion or replacement (i.e. technology refresh) as needed over its operating life.

The Science Data Center also maintains an ongoing applications software development effort, developing ‘science software’ - product generation software embodying science algorithms, e.g. to produce research products to meet the needs of the research team and, in some cases, the overall ESE research program. In some cases the science software would be developed to run in the Science Data Center’s own environment, in other cases the Science Data Center could provide science software to a Backbone Data Center for operational production.

Management - The Science Data Center performs a variety of site-level management functions as well as performing direct management of its functional areas.

Site-level management by the Science Data Center includes planning information technology upgrades / technology refreshes, based on assessments of changing research team needs and availability of new technology. developing and maintaining life cycle data management plans (which address data migrations). It also includes coordinating its internal activities with the mission and science team it supports, and its interaction with the ESE, collaboration among ESE data service providers to support science needs, and any other ESE or broader advisory activities that may be appropriate.

The Science Data Center also participates in ESE / SEEDS management processes, strategic planning, and coordination with other data centers and activities beyond ESE/SEEDS.

Facility / Infrastructure - The Science Data Center provides and maintains a fully furnished and equipped, environmentally controlled, physically secure facility to house its staff, systems, and working storage for its data and information holdings. The Science Data Center provides a separate off-site backup of any new research quality science products generated by the research effort (e.g. that are cited by research team publications), for which the Science Data Center might use the services of a Backbone Data Center.

The Science Data Center ensures system and site security according to established NASA security policies and practices.

The Science Data Center performs resource planning, logistics, supplies inventory and acquisition, and facility management. It provides for purchase of supplies, facility lease and utility costs and other similar overhead costs, hardware maintenance, COTS licenses, etc.

Systematic Measurements Centers

This section describes the generic Systematic Measurements Center type.

Systematic Measurements Center Concept

The following is the concept for NewDISS Multi-Mission Data Centers, herein referred to as Systematic Measurements Centers, from the NewDISS Concept Paper: “A fourth type of data center is the Multi-Mission Data Center. An example of the type of data activity to be carried out by such a data center is the generation of consistent time-series geophysical parameters, an activity exemplified by the current National Oceanic and Atmospheric Administration (NOAA)/NASA Pathfinder Datasets program, which is funded by NASA's ESE and carried out by PIs at various institutions. These efforts will take on more importance in the future, since NASA ESE has the requirement for generating time-series of geophysical parameters, while the EOS mission strategy has evolved so that it is now designed to accommodate technological change. Thus, these efforts will include construction of the long-time scale datasets from more than one NASA (or other) mission. These data centers will need to address network connectivity as part of their on-going activities. Selection of these services will be driven by PI-teaming arrangements, using either NASA-available resources or competitive alternatives."

Systematic Measurements Center Functions

In general, the Systematic Measurements Center is a potentially long lived data management capability implemented to support a particular data synthesis effort by a limited community of users (which will be called its 'synthesis team'). An example of a data synthesis effort would be research into how to cross-calibrate and consistently map measurements made by different missions (perhaps overlapping or consecutive) in order to be able to generate a consistent, continuous, long-term, research quality data set spanning multiple instruments/missions, validation of the cross-calibrated data sets, and then the production of the long time series data set. Such a production effort could be quite intensive in order to accomplish in a reasonable time the generation of a long time series data set involve handling many year's worth of a number of good sized data sets. The synthesis effort could continue adding new data sets to the mix from which its products are produced, extending its time series. The Systematic Measurements Center operates in a research environment, without the need for robustness and performance as would be the case for an operational environment.

The distinction drawn between a Science Data Center and a Systematic Measurements Center is that the former supports a particular research effort, while the latter supports a data synthesis effort that might require an extended period of time to complete, and which would enable future science efforts using the new, research quality, long time series data sets it produces.

Ingest - The Systematic Measurements Center obtains data and products and all supporting documentation needed for its data synthesis effort from a variety of sources, including other ESE data service providers, other agency data centers, etc. The ingest would not be performed on a time critical, operational basis, but could involve large amounts of data if long time series of large data sets are involved.

Processing - The Systematic Measurements Center would perform processing of new data synthesis products (such as long time series data sets) developed by the synthesis team on an ad hoc basis. This processing could be a major effort, for example if the objective is a long time series product produced from a number of large, multi-year input data sets. The Systematic Measurements Center could accomplish a large scale processing effort (such as a major effort to generate a long time series data set once the cross-calibration, mapping, etc., involved had been tested and validated) through a partnership with a Backbone Data Center or other processing facility.

Archive - The Systematic Measurements Center would not perform an archive function per se, but would maintain working storage of data and products obtained from other sources and new data synthesis products generated by the center. This could involve large data volumes, and the working storage would be configured to facilitate the processing effort.

Search and Order - The Systematic Measurements Center provides a search and order capability tailored to meet the needs of the synthesis team it supports, perhaps supplemented to an additional capability for allowing other interested scientists to search for and order certain products the synthesis team deems to be ready for use beyond the immediate work of the science team prior to their availability from a Backbone Data Center.

Access and Distribution - The Systematic Measurements Center generates complete documentation any new data synthesis products developed by the synthesis team that are new research quality products to be made available to the general science community, including full, documentation of the cross-calibration and any other steps taken to build the consistent time series. The Systematic Measurements Center will make the products collected to support the data synthesis effort readily available to members of the synthesis team. The Systematic Measurements Center will also transfer new research quality data synthesis products and documentation to a Backbone Data Center when broader distribution of those products is appropriate, or at the conclusion of the data synthesis effort.

User Support - The Systematic Measurements Center provides close support to members of the synthesis team it supports, including a help desk supplemented by on-line aids (e.g. FAQs).

Instrument / Mission Operations - None.

Sustaining Engineering - The Systematic Measurements Center performs software maintenance as needed.

Engineering Support - The Systematic Measurements Center performs engineering support functions as needed.

Technical Coordination - The Systematic Measurements Center participates in SEEDS system level processes, including coordination on data management, documentation standards, data stewardship (including standards for content of life cycle data management plans), standards and best practices (including quality assurance standards and practices), interfaces, security, common metrics, and interoperability (e.g. for data access and integration), across / within SEEDS and with other systems and networks as needed to support the ESE program.

The Systematic Measurements Center may cooperate with other ESE data service providers in representing ESE / SEEDS in broader community processes in areas such as standards, interoperability, data management, security, etc.

The Systematic Measurements Center also participates in SEEDS level and/or bilateral processes to coordinate production and delivery of products between itself and other ESE data service providers.

Implementation - The Systematic Measurements Center develops the data and information system capabilities it requires by the to perform its mission, including initial design and implementation of the data system (hardware and system software) and applications software and expansion or replacement (i.e. technology refresh) as needed over its operating life.

The Systematic Measurements Center also maintains an ongoing applications software development effort. Applications software can include software to perform data services (e.g. subsetting, reformatting, reprojection, etc.) for more of its products, software tools for use by users to unpack, subset, or otherwise manipulate products provided by the Backbone Data Center, product generation software embodying science algorithms, e.g. to produce a product to meet a particular user need, and to perform a ‘data mining’ or data integration operation to meet a user need.

Management - The Systematic Measurements Center performs a variety of site-level management functions as well as performing direct management of its functional areas.

Site-level management by the Systematic Measurements Center includes planning information technology upgrades / technology refreshes, based on assessments of changing mission or user needs and availability of new technology. It includes developing data stewardship practices, performing data administration with science advice (via the User Advisory Group and other appropriate bodies), developing and maintaining life cycle data management plans (which address data migrations). It also includes coordinating its internal science activities and its interaction with the ESE and broader science community, including collaboration among ESE data service providers to support science needs, and any other ESE or broader advisory activities that may be appropriate.

The Systematic Measurements Center also participates in ESE / SEEDS management processes, strategic planning, and coordination with other data centers and activities beyond ESE/SEEDS.

Facility / Infrastructure - The Systematic Measurements Center provides and maintains a fully furnished and equipped, environmentally controlled, physically secure facility to house its staff, systems, and data and information holdings, including a separate off-site backup facility for its new research quality data synthesis products, for which the Systematic Measurements Center might use the services of a Backbone Data Center.

The Systematic Measurements Center ensures system and site security according to established NASA security policies and practices.

The Systematic Measurements Center performs resource planning, logistics, supplies inventory and acquisition, and facility management. It provides for purchase of supplies, facility lease and utility costs and other similar overhead costs, hardware maintenance, COTS licenses, etc.

Applications Center

This section describes the generic Applications Center.

Applications Center Concept

ESE's Applications Program mission ("Earth Science Enterprise Applications Strategy for 2002-2012", January 2002) is to "Expand and accelerate the realization of societal and economic benefits from Earth science, information, and technology". The overarching goal of the Applications Program is "to bridge the gap between Earth system science research results and the adoption of data and prediction capabilities for reliable and sustained use in decision support". Applications program implementation is seen as selecting applications projects (based on criteria discussed in the strategy document) that would proceed through the steps of applications research, validation and verification, and applications demonstration (depending on their maturity at startup). According to the strategy document, "the desired outcome of applications projects is for the partner organization to use the resulting prototypes, processes, and documentation as benchmarks for operational use. The desired impact is for the application to thrive because the service provider and it customers derive value from the benefits of the operational use of Earth science in serving their decision making processes". At the conclusion of the project NASA would no longer be a source of funding.

Currently, over two hundred applications projects are ongoing (ESE Applications website), organized around several 'themes': resource management (over seventy projects currently), disaster management (over one hundred projects currently), community growth and infrastructure (over twenty projects currently), and environmental assessment (over twenty projects currently). The great majority of these applications projects are focused on developing solutions to specific problems or answers to specific questions within the 'theme' areas, and are not developing a data service provider capability. This prompted much discussion at the February, 2002, SEEDS workshop, which suggested that appropriate 'levels of service' would consider how well problems are solved or questions answered, how flexible the approaches taken were, how extensible the approaches taken would be outside of the specific context in which they were originally developed (e.g. would they work in different geographical regions, different social settings, etc.).

For the purposes of the LOS/CE study, only those current applications projects that can reasonably be seen as having at least some of the attributes of data service providers, will be considered as 'applications centers'. These are a small minority of the over two hundred applications projects ongoing. These will be taken as representative of functionally similar future activities, and information from these will be sought for the comparables database. Three types of applications activities can be viewed as current examples of ‘applications centers’ (although not every individual case within each type may be a good data service provider fit):

Regional Earth Science Applications Centers (RESACs)
The Regional Earth Science Applications Centers are designed to apply remote sensing and attending technologies to well-defined problems and issues of regional significance. There are currently nine public/private consortia throughout the U. S. that form seven RESACs. These consortia will apply state-of-the-art NASA Earth science research results to such diverse areas as precision farm management; monitoring of forest growth and health; regional water resources and hydrology; assessment of the impact of long-term climate variability and change; land cover and land use mapping; agricultural crop disease and infestation detection; management of fire hazards; watershed and coastal management; environmental monitoring; and primary and secondary science education.

Applications Earth Science Information Partners (Type 3 ESIPs)
The Earth Science Information Partners are drawn from academia, government and the private sector. Type 3 ESIPs are charged with developing innovative, practical applications of earth science data for the broader community. Eight Type 3 ESIPs are currently active.

Socio-Economic Data and Applications Center (SEDAC)

SEDAC, the Socioeconomic Data and Applications Center, is one of the Distributed Active Archive Centers (DAACs) in the Earth Observing System Data and Information System (EOSDIS) of the U.S. National Aeronautics and Space Administration. SEDAC focuses on human interactions in the environment. Its mission is to develop and operate applications that support the integration of socioeconomic and Earth science data and to serve as an "Information Gateway" between the Earth and social sciences.

The nine RESACs, eight applications ESIPs, and SEDAC are taken as current examples of Applications Centers, and even within this group there is considerable diversity in the size, scope, function of the activities which range from a NASA-funded DAAC to shared funding partnerships with private groups (universities, private corporations, etc.). SEDAC as a NASA funded EOSDIS DAAC is tightly coupled into EOSDIS and SEEDS processes. Type 3 ESIPs, members along with DAACs in the ESIP Federation, are currently involved in SEEDS processes, and at least while they receive some NASA funding can be expected to continue that involvement, and RESACs logically should be involved with SEEDS in the future. While applications groups receive NASA funding their participation in SEEDS and ESE processes can be supported, but once they become financially independent of NASA their continued participation would become their option.

It is possible to foresee a future in which, given NASA’s commitment to applications expressed in the applications strategy document for 2002 - 2012, there will be a few enduring activities (e.g. SEDAC and some if not all RESACs) and a larger number of activities that arise from cooperative efforts (e.g. applications ESIPs) and go on to independence as they succeed. At any point in time there would be a mixture of enduring activities and other projects at different stages in their evolution, and a much larger number (e.g. about 200 currently) of focused applications projects at various stages in their work.

The intent of this section is to describe an Applications Center as that small subset of the full range of applications activities that have many or at least some of the attributes of a data services provider in that they are not simply targeted on a specific problem for a specific user but offer services or products to meet the needs of a broader community.

Applications Centers will obtain NASA Earth science products and use these, sometimes in conjunction with other Earth science data or any kind of other data to produce special products and/or deliver tailored services to an applications community. The products and services may be oriented around a particular problem or applications area. These communities could include agriculture, fisheries, urban planning, resource management, many etc., which could derive value from NASA Earth science products if they were suitably formatted or packaged (e.g. for use with Geographic Information System (GIS) technology) or used in conjunction with other data to produce new products specifically designed to meet their needs.

Although Applications Center type embraces a variety of possible functional models, the approach being taken to cost estimation, i.e. allowing the user of the cost estimation tool to pick needed functions from a general list, allows cost estimates for individual applications centers to be made. The use of an applications center type template would be less helpful in an individual case, but is expected to provide a reasonable approximation when an overall cost estimate for an ESE level data services architecture alternative is being examined.

Applications Center Functions

In general Applications Centers perform the same functions as other data service provider types, the primary distinction being the nature of their user community and therefore their products and services.

Ingest - The Applications Center obtains data and products required as inputs for its applications products from other ESE data service providers, other agency data centers, etc. In some cases the ingest would be performed on a time critical, operational basis, and in other cases might be on an ad hoc or intermittent basis, and could involve large amounts of data.

Processing - The Applications Center would perform processing of new applications products (such as products for agriculture or fisheries) developed by the Applications Center. This processing could be a major effort if low level data sets of large size are used to generate products on a routine basis. Data integration is likely to be a key processing task for Applications Centers, since use of combinations of ESE science products and a variety of different types of data (socio-economic, etc.) to produce new products is a central function.

Documentation - The Applications Center would generate documentation sufficient to support the current use of its products. Documentation would be written for the applications user, and might also include documentation according to standards in use in applications communities. In some cases, where applications products are to be retained for long term use, the Applications Center ensures that its products are documented to the SEEDS adopted standard for long term archiving, working as necessary with external sources (e.g. other data service providers) to capture all needed information. The Applications Center also ensures produces search metadata, product guides, etc., to SEEDS adopted standards, which would likely include FGDC compliant metadata.

Archive - The Applications Center would not be likely to perform an archive function per se, depending perhaps on the commercial value of its products beyond their first use, but would maintain working storage of data and products obtained from other sources and new applications products generated by the center. This could involve large data volumes, and the working storage would be configured to facilitate the processing effort.

Search and Order - The Applications Center serves a broad user community with a robust and flexible search and order capability that supports user interaction with search and order services and, increasingly, supports automated search and order interaction between software running on a user system and the Applications Center system. The search capability allows an applications user to apply criteria that might include applications parameter(s), spatial-temporal coverage, specific product names, etc., to the metadata describing available products and returning to the user listings supplemented by descriptive information of those product types and instances that meet the criteria. The 'order' capability includes a request/permission step, regardless of how implemented (e.g. manual or automated), where a request for a set of data or product instances, perhaps the results of (or a selected subset of the results of) a search, is processed and accepted or denied.

Applications Center search and order includes providing local user interface and capability and may include providing an interface to a broader based, ESE cross-site search and order capability.

Access and Distribution - The Applications Center may distribute its products to either a very limited user community or a very broad user community, operationally or intermittently and/or on an request basis depending on its particular mission or business plan. A key data service provided by Applications Centers is to make its products readily useable by applications communities, for example by providing them in GIS formats, given the currently widespread and rapidly growing use of GIS tools by many applications groups.

User Support - The Applications Center provides effective user support for a focused or wide range of users depending on its particular mission. Its user support includes assistance provided in direct contact with users by user support staff, e.g. responding to queries, taking of orders, staffing a help desk (i.e., staff awaiting user contacts who can assist in ordering, track and status pending requests, resolve problems, etc.), etc. The Applications Center will increasingly offer more automated user support aids (beginning with on-line documentation, FAQ, etc.) to meet increasing demands on user support with the proliferation of data types, data sources, and tools for users. User support staff should include applications expertise to provide users with assistance in selecting and using data.

The Applications Center may coordinates its user support with other ESE data service providers (e.g. for user referral services). It performs outreach to potential new users, and may participate in coordinated outreach activities with other ESE data service providers.

Instrument / Mission Operations - None.

Sustaining Engineering - The Applications Center performs software maintenance as needed.

Engineering Support - The Applications Center performs engineering support functions as needed.

Technical Coordination - The Applications Center participates in SEEDS system level processes, including coordination on data management, documentation standards, data stewardship (including standards for content of life cycle data management plans), standards and best practices (including quality assurance standards and practices), interfaces, common metrics, and interoperability (e.g. for data access and integration), across / within SEEDS and with other systems and networks as needed to support the ESE program.

The Applications Center participates in ongoing examination of the changing needs of the ESE science and applications program and the consequent impacts on the roles, missions, and services of ESE data service providers.

The Applications Center might participate in coordination of user support guidelines and practices across the network of ESE data service providers and with other data centers as needed to support the ESE science and applications program.

The Applications Center cooperates with other ESE data service providers in representing ESE / SEEDS in broader community processes in areas such as standards, interoperability, data management, security, etc.

The Applications Center also participates in SEEDS level and/or bilateral processes to coordinate access and timely delivery of products its application effort requires from other ESE data service providers.

Implementation - The Applications Center develops the data and information system capabilities it requires by the to perform its mission, including initial design and implementation of the data system (hardware and system software) and applications software and expansion or replacement (i.e. technology refresh) as needed over its operating life.

The Applications Center also maintains an ongoing applications software development effort. Applications software can include software to perform data services (e.g. subsetting, reformatting, reprojection, etc.) for more of its products, software tools for use by users to unpack, subset, or otherwise manipulate products provided by the Applications Center, product generation software embodying science algorithms, e.g. to produce an applications product to meet a particular user need, and to perform a 'data mining' or data integration operation to meet a user need.

Management - The Applications Center performs a variety of site-level management functions as well as performing direct management of its functional areas.

Site-level management by the Applications Center includes planning information technology upgrades / technology refreshes, based on assessments of changing mission or user needs and availability of new technology. Depending on the nature of the applications activity, site-level management might include developing data stewardship practices, performing data administration with appropriate advice (via a User Advisory Group or other appropriate body), developing and maintaining life cycle data management plans (which address data migrations). It may also include coordinating its interaction with the ESE and broader applications and science community, collaboration among ESE data service providers to support applications and science needs, annual Enterprise peer review, and support for its User Advisory Group (which includes representation from the applications, education, etc., communities that it serves) and any other ESE or broader advisory activities that may be appropriate.

The Applications Center may also participate in ESE / SEEDS management processes, strategic planning, and coordination with other data centers and activities beyond ESE/SEEDS.

Facility / Infrastructure - The Applications Center provided and maintains a fully furnished and equipped, environmentally controlled, physically secure facility to house its staff, systems, and data and information holdings, including a separate off-site backup facility for its data and information holdings. The Applications Center ensures system and site security according to appropriate security policies and practices, depending on its nature (e.g. commercial practices for private entities, established NASA security policies and practices for NASA funded entities).

The Applications Center performs resource planning, logistics, supplies inventory and acquisition, and facility management. It provides for purchase of supplies, facility lease and utility costs and other similar overhead costs, hardware maintenance, COTS licenses, etc.

Information Center

This section describes the generic ESE Information Center type.

Information Center Concept

In general the Information Center performs many of the same functions as the Back Bone Data Center, except that the Information Center is concerned with information describing data and products (i.e., one or more types of metadata) rather than the data and products themselves. In general the Information Center will obtain its information from other data service providers, assemble it and make it available to its users, and when its users discover data or products they desire, then help (e.g. by providing links to data service provider websites) those users obtain access to the services of source data service providers.

The addition of this data service provider type was based on discussion at the Formulation Team Retreat, November 7-8, 2001, where 'Echo' was suggested as a possible future instance. The Global Change Master Directory (GCMD) is a currently operational instance of the Information Center type.

Information Center Functions

The paragraphs below will discuss the Information Center role in each of the general data service provider reference model's functional areas.

Ingest - The Information Center performs ingest of one or more metadata types, ranging from product instance (e.g. granule) level inventory metadata streams to overall product type descriptions or service descriptions. In some cases the ingest function may be performed on a time critical, operational basis, e.g. for inventory metadata received from other data service providers to be posted to the Information Center's inventory on an operational basis. In other cases, ingest of product type descriptions (etc.) are received on an ad hoc basis and are infrequently updated. Quality control on incoming metadata is critical if the Information Center's database is to be current with consistent and accurate content.

Processing - The Information Center does not perform this function.

Documentation - The Information Center ensures that is own content is consistent and complete and in conformance with adopted ESE / SEEDS standards, but does not generate or maintain any other documentation.

Archive - The Information Center provides working storage for its database of descriptive information.

Search and Order - The Information Center serves a broad user community with a robust and flexible search and order capability that supports user interaction with search and order services and, increasingly, supports automated search and order interaction between software running on a user system and the Information Center system. The search capability allows a user to apply criteria that might include geophysical parameter(s), spatial-temporal coverage, specific product names, etc., to the metadata describing available data and products and returning to the user listings supplemented by descriptive information of those data or product types that meet the user's criteria and, depending on the level of metadata held by the Information Center, data or product instances that meet specific criteria.

Information Center search and order can include providing local user interface and capability and/or providing an interface to a broader based, ESE cross-site search and order capability.

Access and Distribution - While it provides access and distribution of its own metadata, the Information Center facilitates access to the data and products its metadata describes. This might be in the form of links to source data service provider websites, or the ability to accept a user request for relay to a source data service provider.

User Support - The Information Center provides effective user support for a wide range of users who access its metadata holdings and to the source data service providers who provide the metadata.

Instrument / Mission Operations - None.

Sustaining Engineering - The Information Center performs sustaining engineering, with no or very infrequent interruption of operational capabilities.

Engineering Support - The Information Center performs engineering support functions with no or very infrequent interruption of its operations.

Technical Coordination - The Information Center participates in SEEDS system level processes, including coordination on documentation (especially metadata) standards, standards and best practices, interfaces, security, common metrics, and interoperability across / within SEEDS and with other systems and networks as needed to support the ESE program.

The Information Center participates in ongoing examination of the changing needs of the ESE science and applications program and the consequent impacts on the roles, missions, and services of ESE data service providers.

The Information Center participates in coordination of user support guidelines and practices across the network of ESE data service providers and with other data centers as needed to support the ESE science and applications program.

The Information Center cooperates with other ESE data service providers in representing ESE / SEEDS in broader community processes in areas such as standards, interoperability, data management, security, etc.

Implementation - The Information Center develops the data and information system capabilities it requires by the to perform its mission, including initial design and implementation of the data system (hardware and system software) and applications software and expansion or replacement (i.e. technology refresh) as needed over its operating life.

Management - The Information Center performs a variety of site-level management functions as well as performing direct management of its functional areas.

Site-level management by the Information Center includes planning information technology upgrades / technology refreshes, based on assessments of changing mission or user needs and availability of new technology. It includes performing data administration with science advice (via the User Advisory Group and other appropriate bodies), developing and maintaining a life cycle data management plan covering its information holdings). It also includes coordinating its interaction with the ESE and broader science community, collaboration among ESE data service providers to support science needs, annual Enterprise peer review, and support for its User Advisory Group (which includes representation from the science, applications, education, etc., communities that it serves) and any other ESE or broader advisory activities that may be appropriate.

The Information Center also participates in ESE / SEEDS management processes, strategic planning, and coordination with other data centers and activities beyond ESE/SEEDS.

Facility / Infrastructure - The Information Center provides and maintains a fully furnished and equipped, environmentally controlled, physically secure facility to house its staff, systems, and data and information holdings, including a separate off-site backup facility for its metadata. The Information Center ensures system and site security according to established NASA security policies and practices.

The Information Center performs resource planning, logistics, supplies inventory and acquisition, and facility management. It provides for purchase of supplies, facility lease and utility costs and other similar overhead costs, hardware maintenance, COTS licenses, etc.

The Information Center provides facility / infrastructure support at a level that ensures no or very infrequent interruption of its operations.

Long Term Archive Center

This section describes the generic Long Term Archive Center.

Long Term Archive Center Concept

The report, “Global Change Science Requirements for Long-Term Archiving” (USGCRP, March 1999), of the results of the science panel that met in a workshop held at the National Center for Atmospheric Research (NCAR) in October, 1998, discussed the essential functions and characteristics of a long term archiving program.

In general the Long Term Archive Center performs most if not all of the same functions as the Backbone Data Center, with the additional focus on permanent preservation and archiving of data and products and their documentation, and active support to climate research, etc., that requires reprocessing of and/or access to long time series of data and products. The Long Term Archive Center participates with ESE data service providers in life cycle data management planning and in a process for obtaining science guidance and priorities for long term archiving.

Long term archiving is strictly speaking not an ESE responsibility, but inclusion of a hypothetical Long Term Archive Center type is intended to support planning that NASA is doing with NOAA and USGS, the agencies who have (with NARA) the long term archive responsibility.

Long Term Archive Center Functions

The paragraphs below will discuss the Long Term Archive Center role in each of the general data service provider reference model's functional areas, drawing on the USGCRP report cited in Section 3.7.1 above. Items in the functional discussion below that are explicitly derived from that report are indicated by an appended '(USGCRP)'.

Ingest - The Long Term Archive Center performs ingest of a wide variety of data and product types, ranging from low level data streams to ancillary data to all of the levels of derived products, and their documentation. These products may be new to the center or may be replacements of earlier versions of products already archived by the center.

It is essential that the Long Term Archive Center verify the integrity and quality of data and derived product and associated documentation as it is ingested into the archive (USGCRP).

The ingest would be a transfer from another data service provider, e.g. a Backbone Data Center, according to scenario to be documented in life cycle data management plans. If the transfer is from a research environment (e.g. a Science Data Center) that Long Term Archive Center should proactively reach out to the research source and develop the needed agreements and procedure, assist in planning documentation, etc., (USGCRP). The transfer could be a single bulk delivery, or staged as a series of deliveries over a period of time. The transfer could be by media or network.

Processing - It is essential that the Long Term Archive Center exercise data to produce new products and/or new versions of old products to validate data and product documentation, identify and resolve problems in the data, provide opportunities to scientists within the center to pursue science interests, produce new or updated products that are of value to the science community, provide an opportunity to rethink and reorganize how the data are stored to take into account user access needs as well as accommodate new storage and access technology, and increase data longevity (USGCRP). Typical science processing / reprocessing efforts could include production of long time series of intercalibrated data sets from multiple sources/ sensors to support climate change research.

Processing / reprocessing by the Long Term Archive Center would be on an ad hoc basis, but with tight quality control.

Documentation - It is most essential that the Long Term Archive Center ensure that its data sets and products in the archive are accompanied by complete, comprehensive, and accurate documentation (USGCRP), in accordance with long term archive documentation standard. The center works as necessary with external data sources (e.g. other data service providers) to capture all needed information.

Archive - The Long Term Archive Center provides a very robust archive capability, performing insertion of data into archive storage, and preservation of data, metadata, and documentation within the archive. Preservation and maintenance of data holdings, including ensuring integrity and quality of the data, products, and associated documentation is an essential function of the Long Term Archive Center (USGCRP). Extension of maintenance to include updating of documentation with user comments on the data or product is desirable (USGCRP).

Preservation measures should include quality screening of data entering and exiting the archive, quality screening of archive media, off-site backup with sampling to verify integrity, and accomplishing migrations from one type of media to another. It is essential that the Long Term Archive Center develop and maintain a multi-year data migration plan, and that the center perform integrity checks on archive media between migrations (USGCRP).

Data migrations to new archive technology should be taken as opportunities for processing / reprocessing (USGCRP).

Search and Order - The Long Term Archive Center serves a broad user community with a robust and flexible search and order capability that supports user interaction with search and order services and, increasingly, supports automated search and order interaction between software running on a user system and the Long Term Archive Center system. The search capability allows a user to apply criteria that might include geophysical parameter(s), spatial-temporal coverage, specific product names, etc., to the metadata describing available data and products and returning to the user listings supplemented by descriptive information of those data or product types and instances that meet the criteria. The ‘order’ capability includes a request/permission step, regardless of how implemented (e.g. manual or automated), where a request for a set of data or product instances, perhaps the results of (or a selected subset of the results of) a search, is processed and accepted or denied.

Long Term Archive Center search and order can include providing local user interface and capability and/or providing an interface to a broader based, ESE cross-site search and order capability.

Access and Distribution - The Long Term Archive Center serves a broad user community with a robust search and order and distribution (electronic and media) service, including offering subsetting, reformatting, repackaging in response to user needs. It is essential that the center provide the next and subsequent generation of scientists with appropriate access to, and facilitate their use of, its holdings, where ‘access’ includes a data set / product search and order function, the ability to deliver data and/or products and supporting information (documentation) on suitable media or electronically, and choices of format, user options such as subsetting, that facilitate access and use (USGCRP).

User Support - The Long Term Archive Center provides effective user support (a user support staff knowledgeable about the data and products, willing and able to help users identify, obtain, and use the products the need, including making referrals to other sources of data - USGCRP) for a wide range of users.

Instrument / Mission Operations - None.

Sustaining Engineering - The Long Term Archive Center performs sustaining engineering, with no or very infrequent interruption of operational capabilities.

Engineering Support - The Long Term Archive Center performs engineering support functions with no or very infrequent interruption of its operations.

Technical Coordination - The Long Term Archive Center is participates in SEEDS system level processes, including coordination on data management, documentation standards, data stewardship (including standards for content of life cycle data management plans), standards and best practices (including quality assurance standards and practices), interfaces, common metrics, and interoperability (e.g. for data access and integration), across / within SEEDS and with other systems and networks as needed to support the ESE program.

The Long Term Archive Center participates in coordination of user support guidelines and practices across the network of ESE data service providers and with other data centers as needed to support the ESE science and applications program.

The Long Term Archive Center also participates in multi-lateral and/or bilateral processes to coordinate production and delivery of products between itself and other ESE data service providers.

Implementation - The Long Term Archive Center develops the data and information system capabilities it requires by the to perform its mission, including initial design and implementation of the data system (hardware and system software) and applications software and expansion or replacement (i.e. technology refresh) as needed over its operating life.

The Long Term Archive Center also maintains an ongoing applications software development effort. Applications software can include software to perform data services (e.g. subsetting, reformatting, reprojection, etc.) for more of its products, software tools for use by users to unpack, subset, or otherwise manipulate products provided by the Long Term Archive Center, product generation software embodying science algorithms, e.g. to produce a product to meet a particular user need, and to perform a ‘data mining’ or data integration operation to meet a user need.

Management - The Long Term Archive Center performs a variety of site-level management functions as well as performing direct management of its functional areas. The Long Term Archive Center provides management for its own operation and staff to support its participation in archive related activities. For example, it is essential that the center be actively facilitate the process for deciding which products to include or exclude from, or remove from, the archive (USGCRP). It is essential this process be driven by science priorities and scientific assessments, and that scientists be actively engaged in the process: setting criteria and making decisions (USGCRP). The Long Term Archive Center would participate with the appropriate ESE data service providers in these processes.

Site-level management by the Long Term Archive Center includes planning information technology upgrades / technology refreshes, based on assessments of changing mission or user needs and availability of new technology. It includes developing data stewardship practices, performing data administration with science advice, developing and maintaining life cycle data management plans (which address data migrations). It also includes coordinating its internal science activities and its interaction with the ESE and broader science community, including a visiting scientist program or equivalent, collaboration among ESE data service providers to support science needs, annual Enterprise peer review, and support for its User Advisory Group (which includes representation from the science, applications, education, etc., communities that it serves) and any other ESE or broader advisory activities that may be appropriate.

Facility / Infrastructure - The Long Term Archive Center provides and maintains a fully furnished and equipped, environmentally controlled, physically secure facility to house its staff, systems, and data and information holdings, including a separate off-site backup facility for its data and information holdings. The Long Term Archive Center ensures system and site security according to established agency security policies and practices.

The Long Term Archive Center performs resource planning, logistics, supplies inventory and acquisition, and facility management. It provides for purchase of supplies, facility lease and utility costs and other similar overhead costs, hardware maintenance, COTS licenses, etc.’

The Long Term Archive Center provides facility / infrastructure support at a level that ensures no or very infrequent interruption of its operations.

Allocation of Requirements / LOS to Data Service Provider Types

This section presents the mapping of the general template of data service provider requirements and levels of service presented in Working Paper 5, "Data Service Provider Reference Model - Requirements / Levels of Service" to the ESE data service provider types discussed in Section 3 above. [The term 'template' is used because the requirements contain placeholders for specifics that must be filled in (i.e. choices between alternatives shown, or between possible levels of service, or replacement of placeholders with lists or numerical values) to generate from the template a set of requirements / levels of service that would apply to a specific ESE data service provider, and that would allow a cost estimate for it to be produced.] This mapping would be the basis for separate requirements / levels of service templates for each data service provider type. They in turn become the basis for the projection of estimated costs for new ESE data service providers of each type.

The requirements / levels of service templates will vary from data service provider type to data service provider type. The different types of data service provider will not all perform the same functions, and will not all meet the same requirements. Indeed, where different data service provider types do have a requirement in common, different levels of service are often appropriate for different data service provider types. The objective of the mapping is to show which of the general data service provider requirements apply to each data service provider type, and where applicable, to indicate minimum, recommended, and desirable levels of service for each requirement.

The tables in this section are arranged to allow convenient comparison of how the requirements / levels of service apply to the different data service provider types. The table below is a sample illustrating the format used in the tables below.

Sample Table

Requirement	Levels of Service	BBDC	MDC	SDC	SMC	AC	IC	LTAC
The data service provider shall ingest…. (2.1 a)	1) operational (time-critical) ingest…	M	M			Dep	R	
	2) routine ingest and verification…					Dep	M	R
	3) ingest on a non-operational basis with verification…			R	R	Dep		M
	4) ingest on a non-operational basis.			M	M	Dep		

The first column in each table presents the requirement (with the section containing the requirement in Working Paper 5, "Data Service Provider Reference Model - Requirements / Levels of Service" in parentheses).

The second through ninth columns are an n by 7 matrix of n levels of service vs the seven logical data service provider types. The logical data set provider types are abbreviated:

BBDC - Backbone Data Center

MDC - Mission Data Center

SDC - Science Data Center

SMC - Systematic Measurements Center

AC - Applications Center

IC - Information Center

LTAC - Long Term Archive Center

The second column contains the draft levels of service defined for the requirement, or "none" if there are no levels of service for the requirement. Entries in the next seven columns indicate if, and if so how, each level of service applies to each logical data service provider type. The possible entries for each cell are as follows:

N/a - the requirement does not apply to the logical provider type.

Blank - the requirement applies to the logical provider type, but the level of service does not.

M - the level of service is the minimum required of the provider type.

R - the level of service is recommended for the provider type.

D - the level of service is desired for the provider type.

Dep - for 'depends' - the requirement applies, but there is no predominant level of service for the provider type - real cases could be at any of the levels of service shown.

Y - when there are no levels of service and the requirement applies.

For a given data service provider type, the entries represent the predominant weight. For example, a data service provider type may ingest a number of different data streams, and a particular ingest level of service might apply for each one. What is indicated in this table is the ingest level of service that best characterizes the data service provider type, especially for the purpose of cost estimation. A similar example would be the backbone data service provider type, which might perform ad hoc as well as operational processing; in such a case the requirements / levels of service will reflect the operational processing. In any case, when a cost estimate is being made for an actual data service provider, its specific requirements would be used, so that in the previous example its cost estimate would not reflect operational processing if it's mission did not include any.

The "desired" case can arise for Applications Centers, which may receive NASA/ESE funding only temporarily, or Long Term Archive Centers which are funded by their host agency.

Note that minimum and recommended levels of service may be indicated, or minimum, recommended and desirable levels of service.

There are a few cases where an actual data service provider of given type might not meet a particular requirement contrary to what is indicated in the table. For example, if a Science Data Service Provider provides its data and products to a Backbone Data Center, then the requirement under distribution calling for a data service provider to provide its data, products, and documentation to a Long Term Archive Center would not apply to that data service provider, and a cost estimate for that Science Data Service Provider would reflect that.

As indicated above the mapping in these tables would be used to write a set of requirements / levels of service templates, one for each ESE data service provider type. Each template could then be turned into a high level requirements statement for a specific data service provider of its type by filling the items left as placeholders in the template.

The next several pages contain the requirements / levels of service to data service provider type mappings.

Ingest

Requirement	Levels of Service	BBDC	MDC	SDC	SMC	AC	IC	LTAC
The data service provider shall ingest the following data [ingest data stream table, listing for each data stream: name, source, product types ingested, product type format, products ingested per day of each type, volume ingested per day]. The input data streams should cover all data to be received by the center, e.g. satellite data streams, ancillary data products, processed products generated by other data service providers, etc., based on its ESE mission. (2.1 a)	1) operational (time-critical) ingest with immediate verification of data integrity and quality;	M	M			Dep	R	
	2) routine ingest and verification of data quality and integrity without tight time constraints;					Dep	M	R
	3) ad hoc or intermittent ingest on a non-operational basis with verification of data quality and integrity;			R	R	Dep		M
	4) ad hoc or intermittent ingest on a non-operational basis.			M	M	Dep		

Ingest levels of service can be mixed within a data service provider; i.e. different levels may be appropriate for different data streams.

Ingest requirements for Applications Centers can vary widely from case to case.

Processing

Requirement	Levels of Service	BBDC	MDC	SDC	SMC	AC	IC	LTAC
The data service provider shall generate the following standard products, included required Level 1B products [standard product table, listing for each product type/series: name, format, retention plan, product instances produced per day, volume per day, required input data streams] on a highly reliable, operational basis, either on a routine schedule or on-demand, based on its ESE mission. (2.2 a)	1) standard products shall be generated within 2 days of ingest/availability of required inputs.	D	D	n/a	n/a	Dep	n/a	n/a
	2) standard products shall be generated within 7 days of ingest/availability of required inputs.	R	R	n/a	n/a	Dep	n/a	n/a
	3) standard products shall be generated within 30 days of ingest / availability of required inputs.	M	M	n/a	n/a	Dep	n/a	n/a
The data service provider shall generate the following products [product table, listing for each product type/series: name, format,	1) specific targets for processing adopted on a case by case basis.	Y	n/a	R	D	Dep	n/a	D
	2) general goals for processing.	Y	n/a	M	R	Dep	n/a	R

retention plan, average product instances produced per day, average volume per day, required input data streams] on an ad hoc, non-operational basis. (2.2 b)	3) no goals, purely ad hoc processing.	Y	n/a		M	Dep	n/a	M
The data service provider shall reprocess standard products [standard product table] on an ad hoc basis in response to reprocessing requests. (2.2 c)	1) the capacity for reprocessing shall be 9 times the original processing rate.	D	D			n/a	n/a	R
	2) the capacity for reprocessing shall be 6 times the original processing rate.	R	R			n/a	n/a	M
	3) the capacity for reprocessing shall be 3 times the original processing rate.	M	M	R	R	n/a	n/a	
The data service provider shall reprocess standard products [standard product table, listing for each product a reprocessing interval] according to a reprocessing schedule. (2.2 d)	1) reprocessing shall be performed according to a negotiated reprocessing schedule.	R	R			n/a	n/a	R
	2) reprocessing shall be performed to meet the general goals of a nominal schedule.	M	M	R	R	n/a	n/a	M
	3) reprocessing shall be performed following a nominal schedule on a resource / time available basis.			M	M	n/a	n/a	

Requirement	Levels of Service	BBDC	MDC	SDC	SMC	AC	IC	LTAC
The data service provider shall accept science algorithm software from users for [product list], and perform integration and test of the software, and operational execution of the software to produce products. (2.2 e)	1) the data service provider shall accept standard, research product generation software, and/or data integration and data mining software from users;	D	n/a	n/a	n/a	n/a	n/a	
	2) the data service provider shall accept research product generation software and/or data integration and data mining software from users;	R	n/a	n/a	n/a	n/a	n/a	R
	3) the data service provider shall accept standard and/or research product generation software from users;	M	n/a	n/a	n/a	n/a	n/a	
	4) the data service provider shall accept research product generation software from users;		n/a	n/a	n/a	n/a	n/a	M
	5) the data service provider shall accept standard product generation software from users.		n/a	n/a	n/a	n/a	n/a	
The data services provider shall be capable of cross-calibration of data from multiple sources to produce consistent product time series spanning multiple instruments / platforms. (2.2 f)	None.	D	n/a	n/a	Y	n/a	n/a	D
The data service provider shall provide standard metrics on production to the SEEDS Office. (2.2 g)	None.	Y	Y	Y	Y	n/a	n/a	n/a

The processing level of service can vary for different product generation tasks within a site.

Science Data Center and Systematic Measurements Centers would accept, integrate, test, and execute science software developed by their research teams, but (it is assumed) not from other users.

Documentation

Requirement	Levels of Service	BBDC	MDC	SDC	SMC	AC	IC	LTAC
The data service provider shall generate and provide ESE/SEEDS standard compliant catalog information (metadata, including browse) and documentation describing all data and information held by the data service provider. (2.3 a)	1) data and product holdings (including multiple versions of products and corresponding documentation as needed) documented to the ESE / SEEDS adopted standard for long term archiving, including details of processing algorithms, processing history, many etc.	M	R	R	M		n/a	M
	2) documentation ensured to be sufficient for current use (e.g. product type descriptions, product instance (a.k.a. granule) descriptions including version information, FAQs, 'readme's, web pages with links to metadata, user guides, references to journal articles describing the production or use of the data or product).		M	M		R	n/a	
	3) documentation only as received from product provider.					M	n/a	
The data service provider shall update documentation of data and products with user comments. (2.3 b)	1) data and products routinely updated with user comments.	D					n/a	R
	2) data and products occasionally updated with user comments.	R					n/a	M
	3) data and products rarely updated with user products.	M	R	R	R	R	n/a	
The data service provider shall generate and provide DIF (directory interchange format) documents to the Global Change Master Directory on all products available from the data service provider prior to their re-lease for distribution. (2.3 c)	None	Y	Y	Y	Y	Dep	n/a	Y

Archive

Requirement	Levels of Service	BBDC	MDC	SDC	SMC	AC	IC	LTAC
The data service provider shall add to its archive or working storage the following data and products [archive product table, drawn from ingest data stream table, standard product, and ad hoc product tables and reprocessing volume] and related documentation / metadata. (2.4 a)	None	Y	Y	Y	Y	Dep	n/a	Y
The data service provider shall provide for secure, permanent storage of data at the “raw” sensor level (NASA Level 0 plus appended calibration and geolocation information). (2.4 b)	None	Y	Y	n/a	n/a	n/a	n/a	Y
The data service provider shall provide for secure storage of all standard or other science products it produces until the end of the science mission or until transfer to an approved permanent archive, per the applicable life cycle data management plan (or separate retention plan). (2.4 c)	None	Y	Y	Y	Y	n/a	n/a	n/a
The data service provider shall have the capability to selectively replace archived product instances (single or large sets) with new versions, and to selectively update metadata and documentation (e.g. to update quality flags when a product is validated). (2.4 d)	None	Y	n/a	Y	Y	n/a	Y	Y

The data service provider shall provide for an [archive] [working storage] capacity of [number] TB. (2.4 e)	1) archive capacity is cumulative sum of all data ingested plus all products generated (including allowance for retaining multiple versions of the same product as required to provide needed support to the provider's science or applications community).	M	M	R	R	Dep	n/a	M
	2) archive capacity is limited to a specified threshold.			M	M			
The data service provider shall perform quality screening on data entering the archive (e.g. read after write check when data is written to archive media) and exiting the archive (e.g. track read failures and corrected errors or other indication of media degradation on all reads from archive media). (2.4 f)	1) exit and entry screening.	R	n/a	n/a	n/a	n/a	n/a	M
	2) entry screening.	M	n/a	n/a	n/a	n/a	n/a	
The data service provider shall take steps to ensure the preservation of data in its archive. (2.4 g)	1) 10% per year random screening.	D	n/a	n/a	n/a	n/a	n/a	R
	2) 5% per year random screening.	R						M
	3) 1% per year random screening.	M						
The data service provider shall provide a backup for its [archive] [working storage]. (2.4 h)	1) full off-site backup, with regular sampling to verify integrity.	M	R			n/a		M
	2) partial, [Backup Fraction - % of archive backed up], off-site backup, with sampling.		M	R	R	n/a		
	3) partial, [Backup Fraction - % of archive backed up], on-site backup with sampling.			M	M	n/a	M	
The data service provider shall use robust archive media. (2.4 i)	1) archive media compliant with best commercial practice.	M				n/a	n/a	M
	2) archive media and system vendor independent.		R	R	R	n/a	n/a	
	3) archive media vendor independent.		M	M	M	n/a	n/a	
The data service provider shall plan and perform	1) planned migration.	R	n/a	n/a	R	n/a	n/a	M

shall plan and perform periodic migration of archive to new archive media / technology. (2.4 j)	2) no planned migration, but ad hoc migration as need is seen to arise.	M	n/a	n/a	M	n/a	n/a	
The data service provider shall provide standard metrics on archive to the SEEDS Office. (2.4 k)	None	Y	n/a	n/a	Y	Y	Y	D

Search and Order

Requirement	Levels of Service	BBDC	MDC	SDC	SMC	AC	IC	LTAC
The data service provider shall provide users with access to all metadata and information holdings. (2.5 a)	1) public access to all users.	M					M	M
	2) access to the science and applications community.			R	R	M		
	3) access to a limited team of scientists.		M	M	M			
The data service provider shall provide a world wide web accessible search and order capability to [all users (including the general public) consistent with SEEDS standards and practices][to a limited set of science team members]. (2.5 b)	1) allow search for instances of multiple product types that pertain to a specified object or phenomenon (e.g. a named hurricane, a volcanic eruption, a field campaign, etc.).	D				n/a	D	D
	2) allow search for instances of multiple product types by geophysical parameter(s), time, and space applied across multiple product types.	R				n/a	R	R
	3) allow search for instances of multiple product types by common time and space criteria (coincident search).	M	D	D	D	n/a	M	M
	4) allow search for instances of single product type by time and space criteria.		R	R	R	n/a		
	5) allow search for particular instances of a product type from a list of those available.		M	M	M	n/a		
The data service provider shall provide the user with the option of quickly viewing information describing any product returned as meeting search criteria. (2.5 c)	1) descriptive information includes detailed algorithm and use explanations, references to a few published papers that describe the production or use of the product, standard guide and DIF metadata.	D	D	D	D	n/a	n/a	D

	2) descriptive information includes references to a few published papers that describe the production or use of the product, standard guide and DIF metadata.	R	R	R	R	n/a	n/a	R
	3) descriptive information includes standard guide and DIF metadata.	M	M	M	M	n/a	n/a	M
The data service provider shall provide an interface for system-system search and order access as well as an interface for human users. (2.5 d)	None.	Y	n/a	n/a	n/a	Dep	Y	Y
The data service provider shall provide an interface to and support selected external catalog search capabilities. (2.5 e)	None.	Y	n/a	R	R	Dep	n/a	Y

Search and order requirements and levels of service for the Science Data Center and Systematic Measurements Center that go beyond meeting the needs of the science teams they support apply when these providers retain science quality products and make them more widely available.

Access and Distribution

Requirement	Levels of Service	BBDC	MDC	SDC	SMC	AC	IC	LTAC
The data service provider shall provide users with access to all data and product holdings, including all standard science products (Level 1b, Level 2, and Level 3) produced by the data service provider. (2.6 a)	1) public access to all users.	M				Dep	M	M
	2) access to the science community.			R	R	Dep		
	3) access to a limited team of scientists.		M	M	M	Dep		
The data service provider shall provide data and products to users in (at a minimum) one of the SEEDS core formats. (2.6 b)	None.	Y	Y	Y	Y	n/a	n/a	Y
The data service provider shall enhance its distribution capability with supporting data services such as subsetting, resampling, reformatting (e.g. to GIS formats), reprojection and/or	1) supporting data services available for most archived data and products.	R				Dep	n/a	R
	2) supporting data services available for less than half of archived data and products.	M	R	R	R	Dep	n/a	M

formats), reprojection and/or packaging to meet the needs of its users. (2.6 c)	3) supporting data services available for a few selected data and products only.		M	M	M	Dep	n/a	
The data service provider shall provide data to users on an [operational, subscription, and/or in response to request] basis. (2.6 d)	None.	Y	Y	Y	Y	Y	n/a	Y
The data service provider shall provide an interface for system to system network delivery of data and products. (2.6 e)	None.	Y	R	D	D	D	n/a	Y
The data service provider shall perform timely distribution of data and products to users by network, providing an average distribution volume capacity of [number] TB per day. (2.6 f)	1) availability of a single product for access by user software within ten seconds.	D				Dep	n/a	D
	2) availability of a single product for network delivery within ten seconds.	R	D	D	D	Dep	n/a	R
	3) availability of a single product for network delivery within ten minutes.	M	R	R	R	Dep	n/a	M
	4) availability of a single product for network delivery within twenty four hours.		M	M	M	Dep	n/a	
The data service provider shall perform timely distribution of data and products to users on SEEDS standard media types in response to user requests, providing an average volume capacity of [number] TB per day. (2.6 g)	1) shipping of media product within three days of receipt of request.	R				Dep	n/a	D
	2) shipping of media product within one week of receipt of request.	M	R	R	R	Dep	n/a	R
	3) shipping of media product within one month of receipt of request.		M	M	M	Dep	n/a	M
The data service provider shall have the capacity to distribute products on an average of [number] media units per day. (2.6 h)	None.	Y	Y	Y	Y	Y	n/a	Y
The data service provider with final ESE archive responsibility (i.e., a Backbone Data Center unless, for example, a Science Data Service Provider held its products to the time for their transfer to the long term archive) shall transfer its data, products, and documentation (done to the long term archive	None	Y	Y	Y	Y	n/a	n/a	n/a

standard) to the designated long term archive according to its Life Cycle Data Management Plan. (2.6 i)								
The data service provider shall provide SEEDS standard metrics on distribution to the SEEDS Office. (2.6 j)	None	Y	Y	Y	Y	Y	n/a	n/a

User Support

Requirement	Levels of Service	BBDC	MDC	SDC	SMC	AC	IC	LTAC
The data service provider shall be capable of supporting [number] of distinct, active users per year who request and use data service provider products. (2.7 a)	1) one user support staff member per 100 active users.	R	M	M	R	Dep		
	2) one user support staff member per 500 active users.	M			M	Dep		R
	3) one user support staff member per 1,000 active users.						M	M
The data service provider shall provide a trained user support staff. (2.7 b)	1) below plus science expertise in data / product quality and their research uses.	R	R	R	R	Dep	n/a	R
	2) below plus technical expertise in data structures, use of tools for format conversions, subsetting, analysis, etc.	M				Dep	n/a	M
	3) below plus comprehensive knowledge of details of formats for most if not all products.		M	M	M	Dep	n/a	
	4) user support staff are knowledgeable about the data service provider's holdings and ordering/delivery options.					Dep	n/a	
The data service provider shall provide a help desk function (i.e., staff awaiting user contacts who can assist in ordering, track and status pending requests, resolve problems, etc.). (2.7 c)	1) Help desk staffed seven days per week, twenty-four hours per day.	R				Dep	R	R
	2) Help desk staffed five days per week, twelve hours per day;	M	R	R	R	Dep	M	M

	3) Help desk staffed five days per week, eight hours per day;		M	M	M	Dep		
The data service provider shall provide on-line user support (FAQ, data / product and service descriptions, etc.). (2.7 d)	None	Y	Y	Y	Y	Y	Y	Y
The data service provider shall perform user outreach, education, and training. (2.7 e)	1) Below plus provide user training sessions at universities, schools, etc.	D	n/a	n/a	n/a	Dep	D	D
	2) Below plus expanded booth support including mini-workshops, user training sessions;	R	n/a	n/a	n/a	Dep	R	R
	3) Below plus booth support at four conferences per year;	M	n/a	n/a	n/a	Dep	M	M
	4) Produce and make available outreach material - pamphlets, brochures, posters, etc.		n/a	n/a	n/a	Dep		

Instrument / Mission Operations

Requirement	Levels of Service	BBDC	MDC	SDC	SMC	AC	IC	LTAC
The data service provider shall monitor the status and performance of [name] instruments and in some cases also [name] spacecraft for which it is responsible, generating instrument commands and in some cases space-craft commands as needed. (2.8 a)	None.	n/a	Y	n/a	n/a	n/a	n/a	n/a
The data service provider shall obtain the services of a NASA (or other spacecraft operator as appropriate) mission operations facility to provide instrument and spacecraft data and to receive, validate, and transmit instrument and/or spacecraft commands to the spacecraft. (2.8 b)	None.	n/a	Y	n/a	n/a	n/a	n/a	n/a

Sustaining Engineering

Requirement	Levels of Service	BBDC	MDC	SDC	SMC	AC	IC	LTAC
The data service provider shall maintain and, as needed, enhance custom software it develops to meet its mission needs, and reused software it customizes and integrates, a total of [number] SLOC. (2.9 a)	1) no or very infrequent interruptions of data service provider operations.	R	R				R	R
	2) occasional interruptions in data service provider operations.	M	M	R	R	R	M	M
	3) as needed, with interruptions in data service provider operations a secondary concern.			M	M	M		

Engineering Support

Requirement	Levels of Service	BBDC	MDC	SDC	SMC	AC	IC	LTAC
The data service provider shall perform system administration, network administration, database administration, coordination of hardware maintenance by vendors, and other technical functions as required for performance of its mission. (2.10 a)	1) no or very infrequent interruptions of data service provider operations.	R	R				R	R
	2) occasional interruptions in data service provider operations.	M	M	R	R	R	M	M
	3) as needed, with interruptions in data service provider operations a secondary concern.			M	M	M		
The data service provider shall perform systems engineering, test engineering, configuration management, COTS procurement, installation of COTS upgrades, network / communications engineering and other engineering functions as required for performance of its mission. (2.10 b)	1) no or very infrequent interruptions of data service provider operations.	R	R				R	R
	2) occasional interruptions in data service provider operations.	M	M	R	R	R	M	M
	3) as needed, with interruptions in data service provider operations a secondary concern.			M	M	M		

Technical Coordination

Requirement	Levels of Service	BBDC	MDC	SDC	SMC	AC	IC	LTAC
The data service provider shall provide staff required for participation in SEEDS processes, including ESE data services architecture refinement and evolution, and information technology planning. (2.11 a)	None.	Y	n/a	n/a	n/a	Y	Y	n/a
The data service provider shall provide staff required for participation in SEEDS processes to coordinate data stewardship standards and practices and development and maintenance of standards for content of life cycle data management plans. (2.11 b)	None.	Y	n/a	Y	Y	Y	n/a	Y
The data service provider shall provide staff required for participation in SEEDS processes to coordinate best practices among ESE data service providers, including quality assurance standards and practices for all phases of data services provider functions. (2.11 c)	None.	Y	Y	Y	Y	Y	Y	Y
The data service provider shall provide staff required for participation in SEEDS processes, and cooperating with other ESE data service providers in representing ESE / SEEDS in broader community processes, for developing and maintaining common standards and interface definitions, including those that enable interoperability within the ESE / SEEDS environment and with other systems and networks as needed to support the ESE program. (2.11 d)	None.	Y	Y	Y	Y	Y	Y	Y
The data services provider shall participate in SEEDS level and/or bilateral processes to coordinate production and delivery of products between ESE data service providers. (2.11 e)	None.	Y	Y	Y	Y	Y	Y	Y
The data services provider shall participate in SEEDS processes for coordinating user support guidelines and practices among ESE data services providers. (2.11 f)	None.	Y	n/a	n/a	n/a	Y	Y	Y
The data services provider shall provide staff required for SEEDS coordination of security standards and practices to meet NASA or other established security requirements. (2.11 g)	None.	Y	Y	Y	Y	Y	Y	Y

The data service provider shall provide staff to coordinate standards for common metrics. (2.11 h)	None.	Y	Y	Y	Y	Y	Y	Y
The data service provider shall provide funding for travel to support technical coordination activities. (2.11 i)	None.	Y	Y	Y	Y	Y	Y	Y

Participation by Applications Centers in SEEDS technical coordination would be expected to vary from case to case, depending on the specific mission of each one, and in any given year the degree to which it receives NASA funding. Participation of Applications Centers that become self-sustaining would depend on their view of the benefits of participation.

Long Term Archive Centers are presumed to be funded and operated by other agencies than NASA, i.e. NOAA and USGS. Their participation in SEEDS technical coordination would depend on agreements with NASA and/or their view of the benefits of participation.

Implementation

Requirement	Levels of Service	BBDC	MDC	SDC	SMC	AC	IC	LTAC
The data service provider shall design and a data and information system capable of meeting its mission requirements. The design shall address hardware configuration and interfaces and allocation of function to platform. The design shall address software configuration, including COTS, software re-use, and new custom software to be developed, including science software embodying product generation algorithms and/or software facilitating integration of science software provided by outside source(s). (2.12 a)	None.	Y	Y	Y	Y	Y	Y	Y
The data service provider shall develop a staffing plan that addresses staff required to implement and operate the data service provider over its planned lifetime. The staffing plan shall include a breakdown of positions and skill levels assigned to functions. (2.12 b)	None.	Y	Y	Y	Y	Y	Y	Y
The data service provider shall develop a facility plan, including planning for space, utilities, furnishings, etc., required to	None.	Y	Y	Y	Y	Y	Y	Y

support its staff, data and information system, data storage, etc., and the environmental conditioning to be provided. (2.12 c)								
The data service provider shall accomplish the implementation of its data and information system, including purchase and installation of hardware, purchase or licensing and installation and configuration of COTS software, modification, installation and configuration of re-use software, development of new custom software, and integration of all components into a tested system capable of meeting the data service provider's mission requirements. (2.12.d)	None.	Y	Y	Y	Y	Y	Y	Y
The data service provider shall perform ongoing applications software development. (2.12 e)	1) Below plus implementation of applications software to perform a 'data mining' or data integration operation to meet a user need.	D				D	n/a	D
	2) Below plus implementation of product generation software embodying science algorithms, e.g. to produce a product to meet a particular user need.	R	M	M	M	R		R
	3) Implementation of software tools for use by users to unpack, subset, or otherwise manipulate products provided by the data service provider.	M				M		M
The data service provider shall provide the staff needed to accomplish all needed in-house development and test activities. (2.12 f)	None.	Y	Y	Y	Y	Y	Y	Y

Management

Requirement	Levels of Service	BBD C	MDC	SDC	SMC	AC	IC	LTAC
The data service provider shall provide management and administrative staff to perform supervisory, financial administration, and other administrative functions. (2.13 a)	None	Y	Y	Y	Y	Y	Y	Y
The data service provider shall provide staff required for participation in SEEDS management processes, strategic planning, coordination with other data centers and activities beyond ESE/SEEDS. (2.13 b)	None	Y	Y	Y	Y	D	Y	D
The data service provider shall provide staff with science expertise to coordinate the science activities within the data service provider and its interaction with the ESE and broader science community, including a visiting scientist program (or equivalent) , collaboration among ESE data service providers to support science needs, annual Enterprise peer review, and support for its User Advisory Group and any other advisory activities appropriate given its ESE role and user community. (2.13 c)	None	Y	n/a	n/a	n/a	Dep	n/a	Y
The data service provider shall provide staff with system engineering expertise to plan information technology upgrades / technology refreshes, based on assessments of changing mission or user needs and availability of new technology. (Coordination with other ESE data service providers is included in technical coordination). (2.13 d)	None	Y	Y	Y	Y	Y	Y	Y
The data service provider shall provide staff with data management expertise to develop data stewardship practices, perform data administration with science advice (via the User Advisory Group and other appropriate bodies), develop and maintain life cycle data management plans including data migrations. (Coordination with other ESE data service providers is included in technical coordination). (2.13 e)	None	Y	n/a	n/a	n/a	Dep	n/a	Y

Facility / Infrastructure

Requirement	Levels of Service	BBDC	MDC	SDC	SMC	AC	IC	LTAC
2.14 a: The data service provider shall maintain site, system, and data security according to established NASA or other policies and practices while providing easiest possible access (consistent with required security) to its data and information services for its user community.	None	Y	Y	Y	Y	Y	Y	Y
2.14 b: The data service provider shall provide and maintain a fully furnished and equipped, environmentally con-trolled, physically secure facility to house its staff, systems, and data and information holdings.	None	Y	Y	Y	Y	Y	Y	Y
The data service provider shall provide a backup facility for its data and information holdings. (2.14 c)	1) an environmentally controlled and physically secure off-site backup archive facility.	R	R	D	D			M
	2) an on-site but separate environmentally controlled and physically secure off-site backup facility.	M	M	R	R	R	R	
	3) a backup capability within the data service provider's primary data system(s).			M	M	M	M	
The data service provider shall perform resource planning, logistics, supplies inventory and acquisition, and facility management. (2.14 d)	1) no or very infrequent interruptions of data service provider operations.	M	M	D	D	Dep	M	M
	2) occasional interruptions in data service provider operations.			R	R			
	3) as needed, with interruptions in data service provider operations a secondary concern.			M	M			
The data service provider shall provide network connections and services as needed to support its operations. (2.14 e)	None.	Y	Y	Y	Y	Y	Y	Y

References and Acronym List

The References Section and the Acronym List for all of these Working Papers is in the document "References and Acronyms for the Levels of Service / Cost Estimation Working Papers ".

SEEDS

References and Acronyms
for
Working Papers

Level of Service / Cost Estimation (LOS/CE) Team

April 24, 2002

Introduction

This set of references and list of acronyms accompanies the set of Working Papers prepared by the SEEDS (Strategic Evolution of Earth Science Enterprise Data Systems) Levels of Service (LOS) / Cost Estimation (LOS/CE) study.

The set of working papers that together describe the LOS/CE Study includes the following:

Working Paper 1 - Project Overview and Technical Approach

The first working paper of the set provides an overview of the SEEDS Levels of Service / Cost Estimation Study, a roadmap to the full set of working papers, and a discussion of the technical approach to the requirements analysis and cost estimation phases of the study.

Working Paper 2 - Cost Estimation by Analogy Model
This working paper describes the cost estimation by analogy model that is being developed for this study. This paper will evolve extensively as the work progresses. Its initial focus is on a conceptual description of the model and how it and the cost estimating relationships it uses are expected to develop, scenarios showing how the model will be used, goals and plans for the model prototype, etc.

Working Paper 3 - Data Service Provider Reference Model - Functional Areas
This working paper describes the concepts involved in the Data Service Provider Reference Model, and describes the functional areas / areas of cost comprising the model. The paper reflects the results of the February, 2002, SEEDS Community Workshop, including drawing on material from white papers submitted by workshop attendees.

Working Paper 4 - Data Service Provider Reference Model - Model Parameters
This working paper contains definitions of the parameters that are inputs, outputs, and intermediate parameters used by the cost estimation by analogy model, including those that are elements of the comparables database. It constitutes a data dictionary for the model and database.

Working Paper 5 - Data Service Provider Reference Model - Requirements / Levels of Service
This working paper describes a general set of requirements and levels of service mapped to the functional areas of the Data Service Provider Reference Model. This paper will be maintained and updated as needed through the life of the project. This paper reflects the results of the February, 2002, Community Workshop, draws on white papers submitted by workshop attendees, and includes a new user-oriented view of levels of service.

Working Paper 6 - ESE Logical Data Service Provider Types
This working paper describes an open set of logical ESE data service provider types, each essentially a group of functions clustered around a different type of ESE role or mission as an organizing principle. The paper describes how these logical or conceptual provider types relate to physical entities, e.g. real-world data centers that, given their responsibilities within the ESE program, might embody the functionality of several different provider types. The paper describes how the provider types would be used in ESE architecture studies. The paper reflects the results of the February, 2002, Community Workshop, and draws on white papers submitted by workshop attendees.

Working Paper 7 - Comparables Database
This working paper provides an overview of the Comparables Database, comprising information obtained from existing ESE data activities and other data centers. It includes the database schema or template. It reports on which data centers have provided information to be added to the database, allowing a reader to track the development of the database as the information collection effort proceeds and the paper is updated. The paper does not contain the actual information provided by the sites.

As the initial versions of these working papers are completed they will be made available on the SEEDS website for review and comment, and will be updated in response to feedback and as work on the project progresses.

References

These references were used by one or more of the working papers in the LOS/CE Team set.

1. “NewDISS: A 6-to-10-year Approach to Data Systems and Services for NASA’s Earth Science Enterprise - Draft Version 1.0”, October 2000.

2. “NewDISS Level 0 Requirements”, September 2001, Vanessa Griffin ESDIS/SOO and SEEDS Formulation Team.

3. “ESDIS Project Level 2 Requirements: Volume 5: EOSDIS Version 0, Revision B”, March 2000, GSFC.

4. “ESDIS Data Center Best Practices and Benchmark Report”, September 2001, SGT Inc.

5. “Ensuring the Climate Record from the NPP and NPOESS Meteorological Satellites”, NRC Committee on Earth Studies (CES), September 2000.

6. “Global Change Science Requirements for Long-Term Archiving”, NOAA-NASA and USGCRP Program Office, March 1999.

7. “Survey of Cost Estimation Tools, Final Report” David Torrealba, SGT, March, 2002.

8. “Earth Science Enterprise Applications Strategy for 2002-2012”, NASA/ESE, January 2002.

9. “User Oriented Services Model”, Steve Kempler, Submitted SEEDS Workshop White Paper, February 2002.
10. “SEDAC Inputs to SEEDS Levels of Service Workshop”, Bob Chen, Chris Lenhardt, Submitted SEEDS Workshop White Paper, February 2002.
11. “Operational User Support (OUS) Manifesto”, Hank Wolf, Submitted SEEDS Workshop White Paper, February 2002.
12. “Distributed Data Access, Analysis, and Standards for Earth Science Data”, Menas Kafatos, Submitted SEEDS Workshop White Paper, February 2002.
13. “Outreach, Education Training”, Brenda Jones, Submitted SEEDS Workshop White Paper, February 2002.
14. “Data Management and Services for Global Change Research”, Don Collins, Submitted SEEDS Workshop White Paper, February 2002.
15. “SEEDS: Some Thoughts on Data Management for NASA Missions”, Victor Zlotnicki, Submitted SEEDS Workshop White Paper, February 2002.
16. “Data Services”, Bruce Barkstrom, Submitted SEEDS Workshop White Paper, February 2002.
17. “SEEDS White Paper”, Tom Kalvelage, Submitted SEEDS Workshop White Paper, February 2002.

18. “The Grid: A New Structure for 21st Century Science”, Ian Foster, Physics Today, February 2002

Others TBD.

Acronym List

AC - Applications Center (a logical ESE DSP type)

AO - Announcement of Opportunity

BBDC - Backbone Data Center (a logical ESE DSP type)

CD-ROM - Compact Disk - Read Only Memory

CES - Committee on Earth Studies (National Research Council)

CER - Cost Estimating Relationship

COCOMO - Constructive Cost Model

COTS - Commercial Off-the-Shelf (refers to hardware and software available commercially)

DAAC - Distributed Active Archive Centers (EOSDIS data management / user service elements)

DIF - Directory Interchange Format (used by the GCMD)

DSP - Data Service Provider

DVD - Digital Video (Versatile) Disk

ECS - EOSDIS Core System

EDC - EROS (Earth Resources Observation System) Data Center (USGS, hosts a NASA DAAC)

EDG - EOS Data Gateway

EOS - Earth Observing System

EOSDIS - Earth Observing System Data and Information System

ESDIS - Earth Science Data and Information System (the EOS ground system project at GSFC)

ESE - Earth Science Enterprise (NASA's Earth Science program)

ESIPS - Earth Science Information Partners

FAQ - Frequently Asked Questions

FGDC - Federal Geographic Data Committee

FTE - Full Time Equivalent

FTP - File Transfer Protocol

GCMD - Global Change Master Directory

GIS - Geographic Information System

GSFC - Goddard Space Flight Center (NASA lead center for EOSDIS and SEEDS)

IC - Information Center (a logical ESE DSP type)

LaRC - Langley Research Center (NASA center participating in EOSDIS and SEEDS)

LCDM - Life Cycle Data Management plan

LOS - Level of Service

LOS/CE - Level of Service / Cost Estimation (title of this SEEDS study)

LTA - Long Term Archive

LTAC - Long Term Archive Center (a logical ESE DSP type)

MDC - Mission Data Center (a logical ESE DSP type)

MODAPS - MODIS Adaptive Processing System

MODIS - Moderate Resolution Imaging Spectroradiometer (flown aboard Terra and Aqua)

NASA - National Aeronautics and Space Administration

NCAR - National Center for Atmospheric Research

NewDISS - New Data and Information Systems and Services (now replaced by SEEDS)

NOAA - National Oceanic and Atmospheric Administration

NPOESS - National Polar-orbiting Operational Environmental Satellite System

NPP - NPOESS Preparatory Project

NSIDC - National Snow and Ice Data Center (hosts a NASA DAAC)

PI - Principal Investigator

QA - Quality Assurance

RESAC - Regional Earth Science Applications Center

SCF - Science Computing Facility (operated by EOS PI's)

SDC - Science Data Center (a logical ESE DSP type)

SEDAC - Socio-Economic Data and Applications Center (a DAAC)

SEEDS - Strategic Evolution of ESE Data Systems (new term replacing NewDISS)

SIPS - Science Investigator-led Processing System

SLOC - Source Lines of Code

SGT - Stinger Ghaffarian Technologies (Incorporated), LOS/CE study contractor

SMC - Systematic Measurements Center (a logical ESE DSP type)

SOO - Science Operations Office, within the ESDIS Project

TB - Terabytes

TBD - To be Determined

USGCRP - US Global Change Research Program

USGS - US Geological Survey

Appendix B – Standards for Near-Term and Longer-Term Missions

Near-Term Missions Standards Recommendations

July 30, 2002

SEEDS Near-Term Mission Standard Study Team

Contributors

Richard Ullman, NASA/GSFC, Study Team Lead

Jingli Yang, ERT, Study Team

Cheryl Craig, NCAR, Study Team

John Evans, GST, Study Team

Larry Klein, L-3 Analytics, Study Team

Dorian Shuford, ERT, Study Team

Siri Jodha Singh Khalsa, L-3 Analytics, Study Team

Matt Smith, UAH, Study Team

Table of Contents

List of Tables and Figures

1.0 Introduction

1.1 SEEDS Goals and Strategy

SEEDS, previously called NewDISS, involves the Strategic Evolution of the Earth Science Enterprise Data Systems to serve research and application needs in the next ten years. Its primary goal is to support NASA's Earth Science Enterprise (ESE), which, in turn, contributes to the US Global Change Research Program (USGCRP). As such, SEEDS is driven principally by the objectives of scientific research, but must also serve the needs of both scientific research and a wide variety of practical applications.

Future ESE data systems will consist of a heterogeneous mix of interdependent components derived from the contributions of numerous individuals and institutions. These widely varying participants will be responsible for data management functions including data acquisition and synthesis; access to data and services; and data stewardship.

"An important premise underlying the operation of [the ESE network of data systems and services] is that its various parts should have considerable freedom in the ways in which they implement their functions and capabilities. Implementation will not be centrally developed, nor will the pieces developed be centrally managed. However, every part of [the ESE network] should be configured in such a way that data and information can be readily transferred to any other. This will be achieved primarily through the adoption of common standards and practices [1]."

Figure 1.1.1 is a simplified data flow diagram of the ESE network of data systems and services [1]. Five types of data centers, namely Backbone Processing Centers, PI-managed Mission Data Centers, Science Data Centers, Applications Data Centers, and Multimission Data Centers are shown in the diagram. Several data flows, such as data flows from PI-managed Mission Data Centers to Multimission Data Centers and vice versa, from Science Data Centers to Applications Data Centers and vice versa, from Science Data Centers to Science Data Center, from PI-managed Mission Data Centers to PI-managed Mission Data Centers, etc. are omitted for simplicity. Four different types of data flow are identified in the diagram. Internal data flow refers to data flow inside each data center. L0 or spacecraft data flow refers to spacecraft or level 0 data flow between mission operations, PI-managed Data Centers or Multimission Data Centers, and Backbone or Long-Term Archive Data Centers. Distribution flow denotes data distribution to end-users. System interchange flow denotes data exchange between data centers. As suggested by Figure 1.1.1, the ESE network provides a means for opening numerous new channels for Earth Science satellite data streams to reach the user community. Such data streams will flow to users both directly from mission data processing centers as well as via many intermediate information providers.

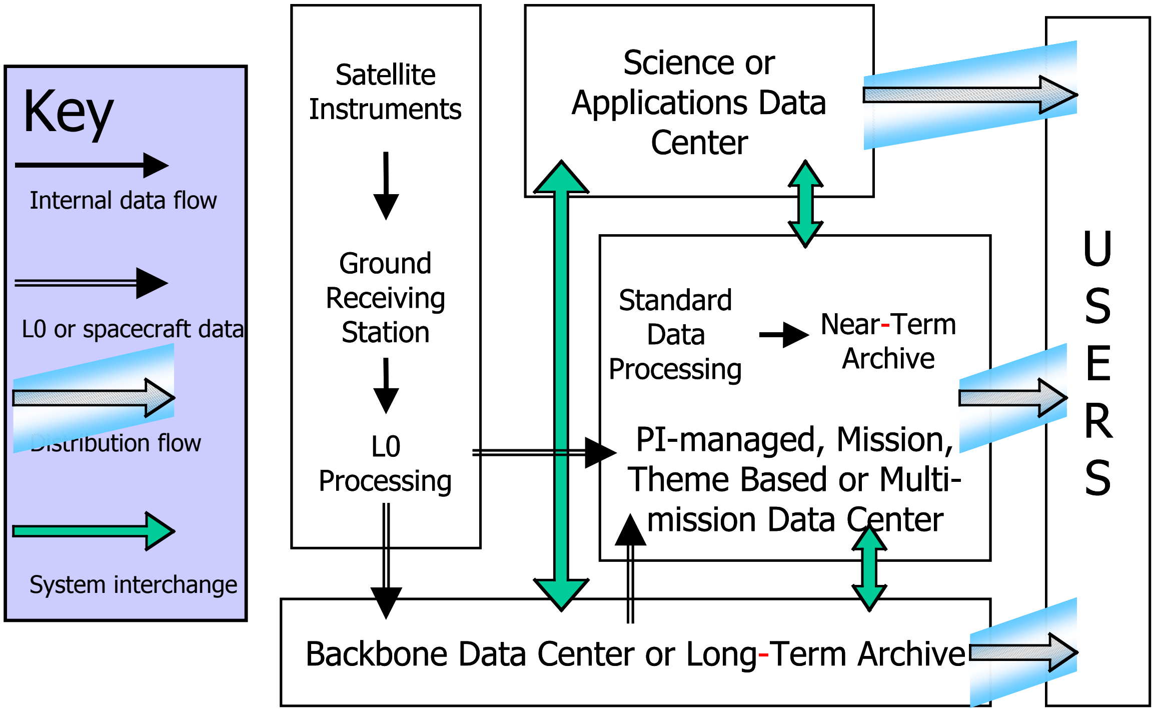


Figure 1.1.1 Simplified ESE network Data Flow (Adopted from Figure C-2 [1])

The SEEDS Near-Term Missions Standards (NTMS) study group is tasked to make recommendations for the use of standards by the ESE near-term missions (described in the Appendix, Section 1.0). These standards are not meant to prescribe the ways that each near-term mission manages data internally or the L0 or spacecraft data flow. Instead, the recommended standards pertain to the data distribution to end-users and to the data interchange between the ESE network of data systems and services components (i.e., between different data centers as shown in Figure 1.1.1).

1.2 The Rationale for Standards

Standards aid in interoperability between data systems and facilitate access by users and the software they use. The successful adoption and use of standards for the ESE network of data systems would reduce the cost and enhance the efficiency of data system development and maintenance. Use of standards for the interchange among the ESE data and service components also makes it easy for data and service providers to join the ESE network of data systems without negotiating one-to-one agreements with each potential provider. The standards that the NTMS study group is addressing include data packaging standards, data service interface standards, metadata standards, and documentation standards, as defined below.

- *Data Packaging Standards* define how to package or encode data that is stored on a computer or transferred from one system to another. Software libraries may be available to facilitate decoding, encoding, or manipulating data packaged in a particular way.

- *Content Standards* for data or metadata define the information elements and their intended meaning (semantics), independently of how these elements may be encoded in files (their syntax). Two or more encodings of the same content standard can be mapped (machine-translated) to each other with no loss of information.
- *Data Service Interface Standards* specify data access requests and service invocations between ESE data and services components, usually over a network. These interface standards are defined independently of the data's packaging (encoding). Web service standards, driven by electronic commerce and other markets, are a particularly promising class of service interface standards in the World Wide Web context.
- *Metadata / Documentation Standards* provide a common lexicon and a set of attributes describing data to ensure that users can 1) find the data in catalogs, registries, and other indexes; 2) interpret the data unambiguously; and 3) apply system services correctly. Metadata is usually highly structured and formalized, whereas, documentation usually refers to more free-text descriptions. Most metadata and documentation standards are content standards (format-independent); XML is a popular encoding for metadata.

For years, various satellite missions and scientific communities have found ways to use each other's data, but stable, rich standards can further promote opportunities in research and applications for data users worldwide. The evolution of these standards over the past 25 years or so has largely been driven by specific science communities with a goal of making life easier for themselves. The past 10 years or so has seen ever wider global scientific communities tied together through the Internet with a goal of still faster-paced data exchange and hopefully faster-paced research results. However, the diversity of available data sources and data standards presents a significant challenge to Earth science researchers, especially interdisciplinary Earth scientists.

As almost any researcher can attest, a substantial portion of the resources required to perform an investigation are expended on locating, obtaining, and then reading and possibly reformatting the necessary data. Standardization of data formats, metadata, and documentation can lower the threshold on data exchange between the ESE network of data systems and services components and the user access to the data products. The Internet offers a compelling example of the essential role standards play in facilitating data exchange. Without the underpinnings of the Internet - TCP/IP, HTML, SMTP, GIF, JPEG, PDF, etc., the explosion of information exchange brought about by the Internet could never have happened.

1.3 Assumptions

This study focuses on near-term missions that are already in formulation and is aimed to provide concrete, specific recommendations for the near-term missions' use. The following assumptions are made to carry out this study.

1. The emerging field of Web Services is driving rapid development of data format-neutral service interface standards. Examples relevant to ESE data include the OpenGIS Web Map Service and Web Coverage Service. However, the use of

online services is still only emerging in practical ESE work; it will take some time before Web Services become a part of mainstream data access and distribution.

2. For the near-term missions, the preferred mode of delivering data remains the transfer of discrete files. Therefore the file format itself is critical to the interchange standard.
3. Content data standards (define the information elements and their intended meaning (semantics), independently of their syntax) provide well-known semantics that can support interoperability through translators or cross-reference tables. The leading definition for such standards is the Federal Geographic Data Committee (FGDC) that has developed *Content Standard for Remote Sensing Swath Data* and *Content Standards for Digital Orthoimagery*. However in practice, content standards alone may not suffice for transferring complex data between different user communities without information loss or distortion.
4. The processes of standards development and adoption are the responsibility of the long-term standards study team.

1.4 Methodology

This document provides recommendations for the use of standards by the near-term missions. We analyzed what standards are currently in use in the near-term heritage missions and other EOS missions, posing questions such as: What are the lessons learned on implementing and using those standards currently in use? What are the lessons learned from other government agencies such as NOAA? What criteria should we use to evaluate different standards? What feedback do data producers and data users have on standards? What standards do users think NASA should use in the future? Once we provide recommendations, how can the recommendations be implemented for the near-term missions? What respective activities should be supported in order to facilitate the adoption of the standards?

This report intends to answer these questions. It is based on a previous report entitled, "Near-Term Missions and Standards Survey," which examines near-term missions and heritages missions and standards in use by the heritage mission data management systems as well as several emerging standards. Most of the content of the survey report are included in the Appendices as background materials. In this report, we present a summary of the heritage missions and standards in use in the heritage missions. We review lessons learned from implementing and using standards in heritage missions and in some NOAA missions. We compare standards based on essential standards concepts. In addition, we develop a suite of standards evaluation criteria and carry out a standards analysis. The results from the standards analysis are presented.

In order to include data users' and data producers' feedback on current data and metadata standards in use in the ESE missions in the study, we conducted a user interview/survey; this report summarizes and analyzes the results from the interview and survey.

References:

[1] A 6 to 10 Year Approach to Data Systems and Services for NASA’s Earth Science Enterprise; Draft Version 1.0; February 2001; Section A.3.

2.0 Near-Term Mission and Heritage Mission Standards

2.1 SEEDS Near-Term Missions

The missions that SEEDS is initially targeted to support include the following eight near-term missions (Table 2.1.1). A detailed description of these missions can be found in the Appendix, Section 1.

Table 2.1.1 SEEDS Near-Term Missions

Mission Name	Phase	Anticipated Launch Date
Landsat Data Continuity Mission (LDCM)	Formulation	2005
NPOESS Preparatory Project (NPP)	Formulation	2005
Ocean Surface Topography Measurement (OSTM)	Formulation	2005
Ocean Vector Winds	Formulation	2006
Global Precipitation Measurement (GPM)	Formulation	2007
Solar Irradiance	Formulation	2007
Carbon Cycle Initiative (CCI)	Pre-Formulation	2008-2012
Total Column Ozone	Pre-Formulation	N/A

See Acronym List if needed

A summary of the near-term mission instruments, data formats, and metadata standards is described in Table 2.1.2. As shown in the table, LDCM, the first near-term mission, has already decided the data and metadata standards they plan to use for the mission data products (specified in the Request For Proposal (RFP) they released October 2001). Our recommendations for the use of data, metadata, and data interfaces in near-term missions may, or may not, impact the LDCM mission.

Table 2.1.2 SEEDS Near-Term Mission Standards

Missions	Instrument	Data Format	Metadata Format
LDCM	Not specified	1. HDF 2. GeoTIFF 3. L7 Fast Format	1. ECS 2. FGDC
NPP	ATMS	N/A	N/A
	CrIS		
	VIIRS		
OSTM (or Jason-2)	N/A	N/A	N/A
Ocean Winds	Seawinds	N/A	N/A
GPM	Dual Frequency Radar (DFR)	N/A	N/A
	Advanced TRMM Microwave Imager (TMI)		
	Nadir-viewing Microwave Radiometer		
Solar Irradiance	N/A	N/A	N/A
CCI Missions: 1. Pathfinder CO_2 2. Ocean Carbon 3. Low Density Biomass	A passive spectrometer	N/A	N/A
	A rotating scanner telescope		
	A hyperspectral imager		
	A P-band SAR and an imaging laser altimeter		

Missions	Instrument	Data Format	Metadata Format
Biomass 4. High Density Biomass 5. Advanced Atmospheric CO_2	A pulsed, dual frequency, tunable laser sounder		
Ozone	Some combination of OMPS-like, TOMS-like, SAGE-like and an IR limb sounder	N/A	N/A

See Acronym List if needed

2.2 Heritage Mission Standards

Data management information for near-term missions and heritage missions is presented in Table 2.2.3.

Table 2.2.3 SEEDS Heritage Missions Data Management Information

Mission	Heritage Mission	Heritage Instrument	Production Site	Archive Site	Data Format	Metadata Format
LDCM	Landsat 1-7	TM ETM+	EDC DAAC	EDC DAAC	1. HDF 4 2. GeoTIFF 3. L7 Fast Format	1. ECS 2. FGDC
NPP	Aqua	AMSU HSB AIRS	GSFC DAAC	GSFC DAAC	HDF-EOS 4	ECS
	Terra	MODIS	GSFC DAAC NSIDC DAAC EDC DAAC	GSFC DAAC NSIDC DAAC EDC DAAC	HDF-EOS 4	ECS
OSTM	Jason-1	Poseidon-2 Radar Altimeter	PO DAAC AVISO	PO DAAC AVISO	Native Binary	Custom
		Jason Microwave Radiometer				
	Topex/ Poseidon	Topex Altimeter		PO DAAC	Native Binary for	Custom

Mission	Heritage Mission	Heritage Instrument	Production Site	Archive Site	Data Format	Metadata Format
		Topex Microwave Radiometer	PO DAAC AVISO	DAAC	Binary for low level products Level 3 product	
Ocean Winds		NSCAT	JPL SeaPAC	PO DAAC	HDF 4	Adapted ECS
	Quikscat	Seawinds	JPL SeaPAC	PO DAAC	HDF 4 BUFR	Adapted ECS
	Adeos-2	Seawinds	JPL SeaPAC	PO DAAC	HDF 4 BUFR	Adapted ECS
GPM	TRMM	TMI	GSFC DAAC	GSFC DAAC	HDF 4	ECS
		VIIRS	GSFC DAAC	GSFC DAAC	HDF 4	ECS
		PR	GSFC DAAC	GSFC DAAC	HDF 4	ECS
		CERES	LaRC SIP	LaRC DAAC	HDF 4	ECS
		LIS	GHRC SIP	GHRC	HDF 4	ECS
Solar Irradiance	SNOE	XPS	LASP	LASP	ASCII	Custom
	UARS SOLSTICE	SOLSTICE SIM	UARS CDHF and GSFC	GSFC DAAC	Native Binary Format	Native SFDU format

Mission	Heritage Mission	Heritage Instrument	Production Site	Archive Site	Data Format	Metadata Format
	ACRIM III	TIM	ACRIM III SIPS	LaRC DAAC	HDF 4	ECS
	SORCE	SIM SOLSTICE XPS TIM	LASP SORCE SIP	GSFC DAAC	HDF-5	ECS
CCI	SeaStar	SeaWiFS	GSFC DAAC	GSFC DAAC	HDF FF	ECS
	Terra	MODIS	GSFC DAAC	GSFC DAAC	HDF-EOS 4	ECS
	Nimbus-7	CZCS	GSFC DAAC	GSFC DAAC	HDF DSP CRTT	Native format
	VCL	MBLA	Raytheon ITSS	EDC DAAC	Unknown	Unknown

Mission	Heritage Mission	Heritage Instrument	Production Site	Archive Site	Data Format	Metadata Format
Total Column Ozone	Nimbus-7 Meteor-4 ADOES Earth Probe QuikTOMS	TOMS	GSFC DAAC	GSFC DAAC	HDF-4	ECS
	AURA	OMI	GSFC DAAC	GSFC DAAC	HDF-4 for Level 0 and 1 HDF-EOS 5 for Level 2 up	ECS

See Acronym List if needed

Several observations can be made from Table 2.2.3:

1. Most of the heritage missions use the Hierarchical Data Format (HDF) or HDF-EOS (Earth Observing System) data formats and the EOSDIS Core System (ECS) metadata format for archiving and distribution. Heritage missions that do not use the HDF or HDF-EOS data formats and the ECS metadata format for product distribution are the Jason-1, Topex/Poseidon, and the Upper Atmospheric Research Satellite (UARS) missions. The Jason-1 and Topex/Poseidon missions are heritage missions to the Ocean Surface Topography Mission. UARS is a heritage mission to the Solar Irradiance mission.

2. Several heritage missions distribute their data products in multiple data and metadata formats. For example, Landsat missions distribute their data products in three different data formats, namely HDF, GeoTIFF, and Fast Format, and two metadata formats, ECS and FGDC (Federal Geographic Data Committee). SeaWinds distributes their data products in HDF and BUFR (Binary Universal Format For Representation of data) format. The HDF format is for distributing research data products by the NASA Jet Propulsion Laboratory (JPL) Distributed Active Archive Center (DAAC), while BUFR format is used to distribute operational data products by NOAA NESDIS (National Environmental Satellite, Data, and Information Service).

- Data distribution formats for heritage missions consist of HDF, HDF-EOS, netCDF, GeoTIFF, Fast Format, BUFR, Binary, and ASCII. Metadata distribution formats for heritage missions include ECS, FGDC, and custom formats. A survey and critique of different data standards and metadata standards can be found in the Appendix, Section 2.0 and Section 3.0, respectively.

3.0 Lessons Learned

This chapter presents lessons learned from past experiences with data and metadata standards used for NASA SEEDS heritage missions and NOAA missions. Some of the lessons learned pertain to past experiences with developing or implementing the standards, and others are related to past experiences with using the standards.

3.1 Lessons Learned on Implementing and Using NASA EOS Standards

3.1.1 Landsat 7

Landsat 7 data products are archived in the HDF format but distributed in three different formats: GeoTIFF, Landsat 7 Fast Format, and HDF. Based on statistics collected by the EDC DAAC [Earth Resources Observation System (EROS) Data Center (EDC) Distributed Active Archive Center (DAAC)] User Services from January 1, 2001, to September 30, 2001, most of the users ordered L-7 data either in Fast Format (46%) or in GeoTIFF (42%). Only 12% of the users ordered L-7 data in HDF format. Of the users who ordered data in HDF format, most were from international ground stations and the data product they ordered was Level 0R. HDF is the only format available for Level 0R. These statistics indicate that:

- User communities welcome multiple distribution data formats. Statistics have shown that users order Landsat 7 data in all three available formats with the majority (88%) of the users choosing GeoTIFF or Fast Format. This indicates that for well-developed satellite mission user communities such as the Landsat data user community, multiple data distribution formats are needed. Different users choose different data formats in their applications.
- Heritage mission data distribution formats play an important role. The reason the majority of the Landsat 7 users choose GeoTIFF or Fast Format may be because the Landsat 7 heritage mission Landsat 5 data products are distributed in Fast Format or GeoTIFF format. Thus, users were already familiar with those two formats. It seems natural that users should choose to use a format they are already familiar with rather than switching to a new data format, such as HDF.
- GeoTIFF data format is gaining popularity among Geographic Information System (GIS) users. Landsat Thematic Mapper (TM) data (Landsat 4-5) products have been distributed in Fast Format since 1984. EDC DAAC began distributing Landsat 5 TM data products in GeoTIFF in recent years. However, based on the statistics collected from January 1 to September 30, 2001, almost half (42%) of the users order Landsat 7 data products in GeoTIFF format. As GeoTIFF format is becoming a popular data format in the GIS user community, EDC DAAC is considering distributing other land remote sensing data, such as ASTER (Advanced Spaceborne Thermal Emission And Reflection Radiometer) data products, in GeoTIFF format in addition to the HDF format.

3.1.2 TERRA

The flagship in NASA's Earth Observing System (EOS), Terra launched on December 18, 1999 and began collecting science data on February 24, 2000. There are five instruments onboard Terra, namely MODIS, ASTER, MISR, CERES, and MOPITT (see Acronym List). The data products from Terra, consisting of a great variety of ocean, atmosphere, and land data sets, are archived and distributed in HDF-EOS format as required by the EOS project. Terra metadata conforms to the ECS data model.

In the early 1990's, NASA's Earth Science Data Information Systems (ESDIS) began evaluating data format standards in preparation for the launches of the EOS satellites. In 1993, after careful consideration of over a dozen different formats, ESDIS chose the Hierarchical Data Format (HDF) for EOS standard data products. During the ECS design phase, it was realized that while HDF was a good format to use for storing data, further standardization would be advantageous. HDF provided little convention for associating spatial and temporal information with the science data itself. To enable additional standardization, the HDF-EOS data format was developed. This format adds mechanisms for storing geo-referencing and temporal information, data organization, and metadata storage.

Terra instrument teams and users have had several problems with implementing and using the HDF-EOS standard and the ECS data model.

- The HDF-EOS Grid and Swath provided a natural structure for the bulk of data taken on Terra and other EOS missions; however, there was no convention for storing individual data values. For example, in the case of one producer, real numbers are stored in 14 bits and 2 additional bits are used for a special purpose rather than using all 16 bits to store the number. The HDF-EOS library can access these data; however, translation and other application tools can have problems. If processing is to be performed on individual words or bits, errors can occur if the user is not cognizant of the storage method.

- There was no convention for packaging both HDF-EOS and HDF objects in the same file. All MODIS (Moderate-Resolution Imaging Spectroradiometer) Level 2 and 3 products are different. Even though they use HDF-EOS structures to store their primary data, many and varied vanilla HDF objects are included in MODIS standard products. MODIS also uses global and local text attributes to store non-ECS metadata rather than dumping it all into the ArchiveMetadata attributes as the HDF-EOS design calls for. This implies that software beyond the HDF-EOS library is required to access the additional attributes.

- Even though HDF-EOS provides a standard for packaging geolocation information, there was no detailed standard for actually calculating this information. For example, some ASTER products are geolocated using a geoid (geodetic coordinates) while others are geolocated using an ellipsoid (geocentric coordinates). This is not a priori obvious to data users.

- HDF-EOS has a steep learning curve. Once that hurdle is overcome, platform independence and common packaging provide convenience in access. However, scientists who are used to flat binary format complain about the complexity of HDF-EOS.

- It was a mistake to try to have one HDF-EOS profile to fit all disciplines. In Terra MODIS case, this leads to unproductive wrangling, an overly broad profile, and poor fit for some (maybe all) disciplines. The lesson learned is to develop strong discipline specific profiles and worry about crossing disciplines later.

- An important lesson learned from Terra s not to impose immature standards such as HDF-EOS. All the following are needed in no less than launch time minus three years:
    - Need an expert base before products are defined.
    - Need tools to verify proper implementation.
    - Need experienced help desk support (and more) and to help with implementation.
- There have been many mismatches between ESDT (Earth Sciences Data Type) and metadata output from MODIS production. This has led to a large number of ingest failures. Quality control on the production end is lacking, and it can be traced to the poor versioning on the MODIS processing system end. There would be no problem if the MODIS processing team acquired their Metadata Configuration Files (MCFs) from installed descriptors at the DAACs. In reality, they modify the MCF locally and then send the changes to ECS. As a result, there can be mismatches between the DAACs installed ESDT and what MODIS is using. This problem has all but disappeared since the MODIS processing team is now using only the official MCFs.

3.1.3 AQUA

AQUA is a NASA Earth Science satellite mission mainly designed to study Earth’s water cycle. AQUA was formerly named EOS PM, signifying its afternoon equatorial crossing time, as opposed to the morning equatorial crossing time for TERRA. Aqua will carry six instruments in a near-polar, low-Earth orbit. The six instruments are the Atmospheric Infrared Sounder (AIRS), the Advanced Microwave Sounding Unit (AMSU-A), the Humidity Sounder of Brazil (HSB), the Advanced Microwave Scanning Radiometer for EOS (AMSR-E), the Moderate-Resolution Imaging Spectroradiometer (MODIS), and the Clouds and the Earth's Radiant Energy System (CERES). The MODIS and CERES instruments are the same as those onboard TERRA launched in 2000. The AQUA mission launched in May 2002.

The data format and metadata standards for the AQUA instrument data are the same as those for TERRA, namely the HDF-EOS and the ECS data model, respectively. Lessons learned from the AIRS instrument team (Evan Manning, AIRS principle developer) and the AMSR-E instrument team (Dawn Conway, University of Alabama in Huntsville, Lead Software Engineer for the AMSR-E Science Team) on implementing the data and metadata standards are summarized below.

1. In general, using the HDF-EOS standards requires a fair amount of “buy-in” and has a steep learning curve. Instrument team developers adapted, but casual users had more trouble. For example, it was relatively easy for an instrument programmer to produce the HDF-EOS files using the simple APID. A lot of end-users, however, are reluctant to accept or “buy into” HDF-EOS because it is new. Both the AIRS and the AMSR-E teams found that HDF-EOS is very easy to use.

2. The HDF-EOS format has adequately supported AIRS and AMSR-E requirements, but:
    - The HDF-EOS should explicitly support field annotations. Without a standard, some developers will add their own annotation to internal HDF objects.
    - The field/attribute distinction is not clear. It seems that a swath attribute is anything that does not have a dimension that is a geolocation dimension. HDF-EOS Swath thinks it's anything with less than 2 dimensions.
3. The documentation for the HDF-EOS is nearly adequate. It could really use some good sample programs. For example, provide examples that actually do something non-trivial, such as check for error conditions.
4. While AMSR-E Lead Science Computing Facility (SCF) found that implementation of the required ECS metadata was simple and straightforward; the AIRS team encountered several problems implementing the ECS data model. In fact, the AMSR-E team found the Science Data Processing (SDP) toolkit unnecessary to complete their tasks. It was noted, however, that the ECS keywords should better relate to keywords used in the GCMD (Global Change Master Directory). Problems that the AIRS team encountered are:
    - The ECS tools for implementing the ECS metadata standards are not easy to use. There are some really tricky parts, like setting “hdfattrname” to "coremetadata.0" or "coremetadata" depending on whether it is embedded metadata or not. The interface is generally confusing.
    - The amount of lead-time for adding an ECS Product Specific Attribute or changing attribute valids, etc. is too long.
    - Documentation for the ECS data model is not adequate.
    - The AIRS team supported ESDIS’s (led by Bob Lutz) attempts to add new valids for ScienceQualityFlag. The failure of those attempts makes it hard for AIRS to support data access as they would prefer to.
5. On a general development note, both teams discovered the importance of regular, consistent communications (telecons, meetings, etc.) between the SCF, SIPS (Science Investigator-lead Processing System), DAAC, and ECS.

3.1.4 AURA

Aura is a NASA mission to study the Earth's ozone, air quality, and climate. This mission is designed exclusively to conduct research on the composition, chemistry, and dynamics of the Earth's upper and lower atmosphere by employing multiple instruments on a single satellite. Aura's chemistry measurements will follow-up on measurements that began with NASA's UARS and will continue the record of satellite ozone data collected from the TOMS (Total Ozone Mapping Spectrometer) missions. The satellite will be launched in June 2003 and will operate for five or more years. The Aura data products will be distributed in HDF-EOS5 format. Aura metadata will conform to the ECS data model.

The HDF file format was designed to be a very flexible format. It is able to store many different types of scientific data in a variety of ways. While this flexibility is an asset to customized data storage, it is not ideal when one is trying to ease sharing of data. As there is so much flexibility, two different developers storing the exact same data can store the data in dramatically different ways. To constrain HDF for use in the EOS community, HDF-EOS was developed.

While HDF-EOS constrains HDF with its POINT, GRID, and SWATH interfaces, it is still possible to create two files that are completely different and require dramatically different readers. Areas of potential mismatch include:

- Organization of data fields and attributes
- Dimension names
- Geolocation names and dimension ordering
- Data field names and dimension ordering
- Units for data fields
- Attribute names, values, and units

When the Aura Data System Working Group (DSWG) reviewed the proposed structure of the Level 2 data files from each instrument, it was discovered that each instrument's data files were, at times, quite different. DSWG agreed that with a little work, it was possible to adopt a uniform set of file format guidelines and that it was advantageous to do so. One of the main advantages of this standard is to allow users the ability to use the same set of tools and I/O routines for any of the Level 2 data from instruments within Aura. At the time of this writing, the "HDF-EOS Aura File Format Guidelines" has been adopted by all of the EOS Aura instrument teams. The guidelines contain detailed, specific information on how to store data. All of the items listed above are specifically addressed. As the launch of Aura has not yet occurred at the time of this writing, the outcome of this endeavor has not been determined, but it is hopeful that by adopting a uniform set of strict guidelines that the benefits will be many. The current guidelines can be found at:

http://www.eos.ucar.edu/hirdls/HDFEOS_Aura_File_Format_Guidelines.doc (Microsoft Word version)

http://www.eos.ucar.edu/hirdls/HDFEOS_Aura_File_Format_Guidelines.pdf (Adobe Acrobat format)

3.1.5 QuikSCAT/SeaWinds

The SeaWinds instrument on the QuickScat satellite is a specialized microwave radar that measures near-surface wind speed and direction under all weather conditions and cloud cover. It was launched in 1999 as a follow-on mission to the NASA scatterometer (NSCAT) that flew on the Japanese ADEOS-1 (Advanced Earth Observing Satellite) platform during 1996-1997; and the Seasat-A scatterometer system (SASS), which flew in 1978.

A unique feature of the QuikSCAT/SeaWinds mission is that SeaWinds data are processed, archived, and distributed at both NASA JPL and NOAA NESDIS. SeaWinds data are downloaded from QuikSCAT once every orbit (101 minutes). The stream passes on from the receiving ground station to the Central Standard Autonomous File Server (C-SAFS) at Goddard Space Flight Center (GSFC). The data are then forwarded to both JPL and NOAA. JPL uses these data to produce its science-level wind product, while NOAA uses an altered version of JPL's processing to produce its own Near Real Time (NRT) wind product. This dichotomy can be summarized as follows:

- While the processing software used at NASA JPL and at NOAA NESDIS is the same, data products produced at JPL are research products (with higher accuracy) used for research and in the application community, while data products from NOAA are near real-time products (within 3 hours of observation) targeted for operational users such as the National Weather Services (NWS).

- The SeaWinds products distributed by JPL are in HDF format while data products distributed by NOAA NESDIS are in BUFR format. This is because many operational and modeling users use the WMO (World Meteorological Organization) data standards, BUFR and GRiB (GRidded Binary). NOAA is required to provide data to their operational users in BUFR/GRiB format. For the future, the current plan is to move the NRT processing from NOAA to the Physical Oceanography (PO) DAAC at JPL, starting with the ADEOS-II mission in 2002.

3.1.6 ACRIM

For ACRIM, using HDF-EOS was required; however, since mapping the terrain of the Earth was not necessary (ACRIM is solar pointing), the EOS part did not apply. ACRIM was actually using something akin to a subset of HDF. Because ACRIM used HDF in a limited fashion, enough tools were available, but it still required the team to learn almost everything about HDF in order to determine what functions they actually needed. Overall, HDF was relatively easy to implement. Some lessons learned indicate that the following would have been helpful in the implementation of HDF:

- An instruction manual – "What would have been helpful is a manual with step-by-step instructions; it could have been a quicker implementation."
- Help desk – "Having someone who could spend a little time over the phone would have been very helpful."
- Rectifying the problems with creating HDF files with REAL and INTEGER values.

[Frank Boecherer, ACRIM Science Computing Facility, Personal Communication, June 2002]

3.1.7 SeaWiFS

Ten years ago, when SeaWiFS was in development, HDF had some capabilities that were not supported at that time. In the beginning, HDF was largely an image format; it only supported a limited number of data sets, and it had floating point numbers only. The

SeaWiFS team identified these deficiencies early on; documented and issued reports; then received responses from National Center for Supercomputing Applications (NCSA). As a result, HDF was made more friendly and easier to use. In addition, the parallel development of HDF for use with IDL allowed users to write their own HDF tools. The main thing that was learned through the experience of implementing HDF into the SeaWiFS project was that good user support is essential. The group at NCSA responded to all of their needs at the time. “That was the thing that made it work – user support, help desk.” [Fred Patt, SAIC Project Manager, Personal Communication, June 2002]

SeaDAS (SeaWiFS Data Analysis System) is a comprehensive image analysis package for the processing, display, analysis, and quality control of all SeaWiFS data products, ADEOS / OCTS (Advanced Earth Observing Satellite / Ocean Color and Temperature Scanner, Japan), MOS (Modular Optoelectronic Scanner, Germany), CZCS (Coastal Zone Color Scanner, NASA), and Ancillary data (Meteorological, Ozone). HDF facilitated the development of this powerful tool. The versatility of HDF also allows individuals to develop their own uses within the SeaDAS system. HDF was mandated for the SeaWiFS project because EOS was still under development, and SeaWiFS was to pave the way for future missions. One lesson learned is: allow time to develop tools (or preferably use existing tools) to facilitate ease of use. [Jim Acker, DAAC User Support, Personal Communication, June 2002]

3.1.8 Jason-1

For Jason-1, binary was chosen as the primary data product for historical reasons (continuity). The main advantage of using binary is that it is fast and simple. Once given the read program, it is self-contained. A disadvantage to binary is that each data set requires its own read program.

Initially, one of the problems with HDF was that software to read the format was not widely available, and it did not work on many important computer classes. A second problem, in the past, was that installing the HDF libraries required major system administration knowledge. Also, the initial jump into HDF is difficult and requires a lot of “handholding”, but only for first-time users. However, the beauty of HDF is uniformity across mission data sets.

From these ideas, the main lessons drawn are:

- Before declaring a format "STD", make sure it installs properly and runs on the main machines intended.
- Understand which classes of users will be EXCLUDED by the new format (for example, the simple binary format of Topex can be read on even a windows 95 computer, but HDF will not install there). It is acceptable to exclude classes of users CONSCIOUSLY, but not because of oversight.
- Do not underestimate the “handholding” that will be needed to help users install, then run, the new software. HDF, etc. are not 'read programs,’ they compare to major operating systems or major commercial packages (IDL, Matlab, Mathematica, etc) in their complexity and their installation can be as complex.

[Victor Zlotnicki, Jet Propulsion Laboratory, Personal Communication, June 2002]

3.1.9 AVHRR

AVHRR data format was based on TIROS data for continuity (level 1B, native binary). However, about 2-years ago, NOAA began offering AMSU data in HDF-EOS along with the BUFR and 1B products. The response to HDF-EOS was great. Almost all of the climate scientists are now using the HDF-EOS format by their own choice. In the future, NOAA hopes to offer AVHRR as an HDF-EOS product, due to customer demand. [Ingrid Guch, National Environmental Satellite, Data and Information Services (NESDIS), Personal Communication, June 2002]

The HDF format has already been chosen for the reprocessing of all AVHRR data for JPL. It was known that the data files would need to be compressed, but the problem was, if just a small part of a big data set was needed, the entire file would have to be decompressed and then the small subset would have to be extracted. With HDF, a chunking process exists (also called tiling). This compresses the data in such a way that it allows storage of data sets in chunks that can be decompressed separately. Thus, HDF-4 was chosen for the reprocessing of the AVHRR data. [Peter Cornillon, University of Rhode Island, Oceanography Department, Personal Communication, June 2002]

3.2 Lessons Learned on Implementing and Using other Standards

3.2.1 NOAA Standards

The National Oceanic and Atmospheric Administration's (NOAA's) National Environmental Satellite, Data, and Information Service (NESDIS) operates NOAA’s environmental (weather) satellites and manages the processing and distribution of the data and images these satellites produce daily. NOAA’s operational weather satellite system is composed of two types of satellites: Geostationary Operational Environmental Satellites (GOES) for "now-casting" and short-range warning and Polar-Orbiting Environmental Satellites (POES) for longer-term forecasting. Both types of satellites are necessary for providing a complete global weather monitoring system. The primary customer is NOAA's National Weather Service (NWS), which uses satellite data to create forecasts for the public, television, radio, and weather advisory services.

NOAA NESDIS does not use consistent data and metadata formats for their POES and GOES satellite data archive and distribution. The POES and GOES data are processed by the Information Processing Division (IPD) of the NESDIS Office of Satellite Data Processing and Distribution (OSDPD). The IPD is responsible for ingest, processing, and dissemination of environmental satellite data. The GOES data are distributed in McIDAS formats. The POES weather and climate data products are distributed in various different data formats including flat binary file, Level 1b, GIF, ASCII, BUFR, GRiB, HDF-EOS, netCDF, and McIDAS [1].

- In general, NOAA NESDIS uses multiple distribution data formats to satisfy different user communities’ needs [Ingrid Guch, NOAA NESDIS, personal communication]. The National Weather Service or the modeling community (US and international) uses the WMO data standards, BUFR and GRiB. These users have been relying on NOAA to format the data in BUFR and GRiB (as opposed to them taking the data and running their own converter). The BUFR/GRiB formats are very complex, though, and not generally used by the people outside the modeling community.

- The imaging, climate, and scientific community as well as the NOAA NESDIS maintenance personnel greatly prefer the HDF-EOS data (ease in visualization, combining datasets, using commercial software, etc.). The netCDF format has the same benefit.
- Other experienced users (education, academic, etc.) seem to prefer a binary or ASCII flat file so they can easily manipulate it and add GIS or whatever extensions they like.
- Browsing users (education, some academic folks, etc.) prefer the option of ASCII, spreadsheet, and GIF.
- For satellite data (sensor counts with navigation and calibration appended but not applied), users seem satisfied with the current packed binary file (Level 1b format). The internal NESDIS maintenance personnel have been using an unpacked binary file (Level 1b star) for ease of use in real-time processing. However, this requires recreation of the "unpacked" file from archived metadata and the 1b if reprocessing is necessary (problems occurred in the real-time processing).

Long-term environmental satellite data products are archived and distributed at the NOAA National Climatic Data Center (NCDC). Archive formats used in NCDC are different for different data products. Many products are archived in a custom format and others are in HDF-EOS, Level 1b, ASCII, or JPEG [Kathy Kidwell, NOAA NCDC, personal communication, 2002]. Data distribution formats are the same as the archive formats in NCDC. Lessons learned on NOAA data standards are summarized below:

- Since NOAA is an operational agency and its main customer is the NWS, NOAA NESDIS is required to distribute their satellite data in BUFR/GRiB format to the NWS or the modeling users, although there are many problems with the BUFR/GRiB format [Ingrid Guch, NOAA NESDIS, personal communication; 2002].
- NOAA NCDC has many legacy systems and they have problems translating data to/from BUFR/GRiB format [Geoffery Goodrum, NOAA NCDC, personal communication, 2002].
- The NOAA NESDIS staff have had a positive experience with the HDF-EOS data format [2] and their users, mainly imaging, climate, and scientific communities, like the HDF-EOS format because of the flexibility, tools, and vendor support [3].

3.2.2 The Spatial Data Transfer Standard (SDTS)

The Spatial Data Transfer Standard became a Federal Information Processing Standard (FIPS 173) in 1992, after a 10-year development effort. It was to serve as the national spatial data transfer mechanism for all U. S. Federal agencies, and to be available for use by state and local government entities, the private sector, and research organizations. SDTS specifies exchange format constructs, addressing structure, and content, for spatially-referenced vector and raster data, to facilitate data transfer between dissimilar spatial database systems.[4] The Spatial Data Transfer Standard (SDTS) doesn't prescribe a single data model; rather it provides a set of rules intended to represent virtually any data model.

However, SDTS fell short of its ambitious goals; and the marketplace was slow to accept and support it. Arctur *et al.* [5] list a number of reasons for this:

- *Complexity* - SDTS was driven primarily by large national-level data producers and their needs (very large databases, complex interdependencies, high precision, flexible models, extensive metadata, collaborative updates, etc.). These needs far exceeded those of casual "desktop GIS" users and of most commercial, regional, or local GIS projects, and they stretch even today's GIS technology to its limits. Many people in the GIS community found SDTS to be overly complex, few understood its intended purpose, and thus few chose it when other, more established formats were available.[6] (Arctur *et al.* [5] suggest that as GIS users become more sophisticated, they may demand more of their technology (including data models and formats), and be more able and willing to cope with the implied complexity.)
- *Slow development of the standard in a fast-changing market* - In the decade that elapsed between the first work on SDTS and its final adoption as a standard, the GIS industry grew significantly, and several vendor-specific exchange formats came into widespread use, which satisfied many users' immediate needs, and thus limited the community's interest in using SDTS (which many perceived as yet another format). Even though the standard was mandated for all federal agencies, most data suppliers, responding to user demand, offered alternative data encodings – and only the most curious and experimental users chose SDTS.
- *Limited vendor support* - SDTS got caught in a "chicken-and-egg" situation with GIS vendors: in order to build market demand for SDTS-aware software, data providers needed to produce large volumes of SDTS data. But they needed to use commercial GIS products to build these data; so they had to persuade vendors to produce SDTS products in the absence of customer demand. A few vendors did include STDS conversion tools in their products (e.g., ESRI's Arc/Info, Laser-Scan's Gothic); however different products interpreted SDTS ambiguities differently (see below), so they would often fail to translate unexpected STDS constructs introduced by another vendor's product.
- *Slow development of practical profile*s - SDTS was a very general standard: any practical use of it required users to agree on a particular profile. But due to the complexity of SDTS, and the limited educational material (such as usage examples) available to the geospatial community, it took another four years to complete the first usable profile of SDTS (the Topological Vector Profile). The lack of interest in, and understanding of, SDTS among the GIS community also reduced the demand for useful profiles, and the community's enthusiasm for working on them. In the end, this first profile proved to be both limiting (encoding fairly mundane examples required awkward workarounds) and

unnecessarily complex (it required arc/node/polygon topology, which was unnecessary or even meaningless for many commonly-used cases). [7]

- *Harmonization delays* - Subsequent efforts to define other SDTS profiles (the Raster Profile and Transportation Network Profile) were almost complete when they became mired in attempts to harmonize them with similar standards being developed in NIMA, NATO, and the European Union. This resulted in further delays to their development. (Arctur *et al.* [5] suggest that early harmonization is easier, and that profiles should not be developed so quickly as to overlook other, related standards.)
- *Ambiguity* in the data model (e.g., the cardinality of relationships) and the data semantics (e.g., the meaning of relationships among entities) of SDTS and its profiles limited the utility of SDTS for reliable information transfer. (Arctur [8] likens an SDTS profile to a game in which teams agree on the size of the ball and the shape of the field, but not on the rules of play.) SDTS was supposed to be very general, and to make datasets *self-describing*; that is, the data model could be determined from the dataset contents. But this proved an elusive goal; and thus many even of those who were willing to be “SDTS pioneers” ultimately concluded that its practical value was limited.

In addition, during and after the development of SDTS, new, unanticipated technical expectations arose, which demanded significant technical (re)design and international coordination, and further weakened the community’s support for SDTS:

- a standard means of representing subtiles within a dataset;
- support for permanent, universally unique object identifiers across all datasets;
- support for value-added extensions and incremental updates by users;
- support for tracking changes and historical lineage of features and spatial primitives;
- harmonizing the metadata content with emerging international standards; and
- harmonizing repository organization with emerging OpenGIS software interfaces.

Some of these issues might have been anticipated in the design of SDTS, while others stemmed from the increasing sophistication of GIS products and their users over the years.

The need for harmonization with OpenGIS led to OpenGIS’ work on interface specifications for access to geospatial data (features, coverages, identifiers, etc.). Since the late 1990s, OpenGIS has been the locus of much subsequent work in this area. It focused first on accessing geospatial data (e.g., Simple Features Access for SQL, COM, and CORBA), then on encoding geospatial features in XML (Geography Markup Language (GML)) for transfer between clients and servers.

In summary, the SDTS experience illustrates the importance of keeping pace with technology and market trends and emerging expectations, even after capturing initial requirements. It shows the role of timing: a standard may be “ahead of its time” (arriving before people are ready to understand them or accept more complexity) or “overcome by events" (arriving after people are used to making do without flexible, general, or vendor-independent solutions). Paradoxically perhaps, SDTS was both!

The SDTS experience also underscores the need to balance advanced needs with more basic ones; the importance of good documentation and usage examples; the challenge of "priming the pump" among vendors in advance of market demand; the benefits and risks of harmonizing with related standards; and the futility of mandating a standard that fails to meet a need.

References:

[1] NESDIS Satellite Product Overview Display, http://osdacces.nesdis.noaa.gov:8081/satprod/products/prod_frameset.cfm?prodid=-1

[2] Huan Meng, Doug Moor, Limin Zhao, Ralph Ferraro, HDF-EOS at NOAA/NESDIS; Presentation; HDF-EOS Workshop, 2000, Landover, MD: http://hdfeos.gsfc.nasa.gov/hdfeos/WSfour/meng/hdfeos4.ppt

[3] Andrew S. Jones and Thomas H. Vonder Haar; A Dynamic Parallel Data-Computing Environment for Cross-sensor Satellite Data Merger and Scientific Analysis; Accepted for publication by Journal of Atmospheric and Oceanic Technology; March 2002.

[4] Fegeas, Robin (1995). Q3.3: What is this SDTS thing and is it available via ftp? In *GIS-L / comp.infosystems.gis Frequently Asked Questions and General Info List* (Lisa Nyman, ed.). http://www.prenhall.com/startgis/faq.html.

[5] Arctur, D., Hair, D., Timson, G., Martin, E., and Fegeas, R. (1998) Issues and Prospects for the Next Generation of the Spatial Data Transfer Standard (SDTS). *International Journal of Geographical Information Science* 12(4): 403-425.

[6] Hastings et al. (1996) *The Spatial Data Transfer Standard: Closing the Loop?* - Panel Discussion at GIS/LIS--Denver, Colorado--November 19, 1996. http://www.ngdc.noaa.gov/seg/tools/sdts/gislis_main.html

[7] Kelley, C., and Gosinski, T., 1994: Spatial Data Transfer Standard: do you fit the profile? *GIS World*, August 1994, pp. 48-50. http://wwwsgi.ursus.maine.edu/gisweb/spatdb/urisa/ur94076.html

[8] Arctur, David K., 1996: Spatial Data Transfer Standard: A GIS Vendor's Perspective. Online article. http://web.archive.org/web/19990222153931/www.lsl.co.uk/~arctur/portfolio/sdts.html

4.0 Essential Standards Concepts

Before evaluating individual data or metadata standards, it may be useful to review several key concepts crucial to understanding and comparing standards.

- A comparison with private, ad-hoc, binary information transfer
- Mandatory vs. optional elements of a standard; profiles and extensions
- Abstract vs. implementation standards
- Content and format standards vs. behavior and interface standards

4.1 A Comparison with Private, Ad-hoc, Binary Information Transfer

Webster's dictionary defines a standard as: "That which is established as a rule or model by authority, custom, or general consent." Thus, standards exist only within a community of people sharing certain usage patterns ("custom") or organizational structures (formal "authority" or informal "general consent"). The emphasis in this study is on standards accepted by a fairly broad set of users, publicly documented, and either stable (unchanging) or changeable only by a consensus among these users.

Another aspect of standards is that they govern only a part of the information transfer process. For instance, GeoTIFF codifies the georeferencing of an image, but is silent on the meaning of its pixel values. Whatever a standard does not specify is left to the private (often implicit) understanding of each user community or to ad-hoc ancillary information (such as a README file or a telephone message describing data details). So, at one extreme, complex and rigid standards specify every aspect of information transfer, and at the other extreme, private agreements or ad-hoc communications leave everything implicit or unstructured. Most standards fall somewhere in between; they govern a certain piece of the information transfer process to let a certain set of users communicate or work together, but users must also rely on other standards, private agreements, or ad-hoc qualifiers.

Many earth science data users favor a "raw binary" data format that is both simple and comprehensive. In fact, "raw binary" doesn't actually mean mysterious data files that one must guess at – but, rather, a simple format (often some kind of raster grid) used by a small set of colleagues with little attention to documentation or stability. Thus, "raw binary" denotes not a single format, but as many different formats as there are workgroups. Each such format has a syntax and semantics invented just for that data set or data series, usually without a lot of attention to other related formats, and with many details left implicit or provided "out of band" in a mission report or some other natural-language document.

Such a format may serve many people's immediate needs, for several reasons.

- A given science team often works with only one kind of data and so gets used to the one syntax for that data (*e.g.*, keeps reusing the same parser) and the one set of semantics (*e.g.*, pins a page from the mission report to the cubicle wall).

- Science teams traditionally put a low emphasis on making their data accessible to others outside of their immediate colleagues. They may feel that they have done their job by distributing a simple "README" file with the data.
- ESE data is commonly encoded as raster grids for which one can make easy guesses as to syntax (band-sequential, etc.).
- Most importantly, perhaps, the semantics of much ESE data (platform orientation, sensor model, calibration information, interpretation algorithms) are so complex that bundling them with the data is often difficult; so they tend to remain in a mission report document – or in people's heads. (For example, each MODIS L1b granule has dozens of ancillary data items required for proper interpretation along with several grids of data error and reliability estimates.)

However, the use of raw binary data relies much more on private agreements among colleagues than on documented, consensus standards. It has many of the properties opposite to those of standards, as listed above in the "Rationale for Standards" paragraph. It limits the ability of science teams to move beyond traditional work methods towards more effective interdisciplinary research, collaborative work, and applications. The essential points are:

- Data in a standard formats (should) convey something about the data that its users need to know; whereas, users of binary data must rely on inside knowledge or educated guesses to read and interpret the data.
- Data in a standard format may be used outside of an "inner circle" of colleagues, but only holders of the necessary private information can use raw binary data.
- Standard data formats limit the need for pair-wise translators to and from every possible format, whereas, each raw binary format needs a different translator.
- Standard data formats, by fixing the syntax and semantics of information, allow the possibility of machine-to-machine communication between different systems (that is, interoperability). In contrast, raw binary data requires human inspection an intervention, thus, hindering (preventing) system interoperability.
- Standard data formats facilitate unambiguous transfer of information between users of different systems working with different datasets. This is more difficult with raw binary data, which often loses all but raw pixel values in translation.

In summary, the use of private agreements does not constitute a standard, and so “raw binary” data formats cannot be compared alongside open consensus standards.

Of course, members of a community may choose to turn a private, internal convention into a standard for a wider community by documenting and publishing their shared syntax and semantics and by sticking to what they document (that is, submitting any changes to a public consensus process or formal authority). Most standards are born this way when a usage community publishes its internal conventions to facilitate collaboration with others.

4.2 Mandatory vs. Optional Elements, Profiles and Extensions

Many standards have a set of mandatory elements to ensure basic interoperability plus a set of optional elements to serve a diversity of users and uses. This provides a "base" standard from which a particular community of users may define a *profile* (a more specific standard) to support richer communication among themselves, or more fine-grained control of each other's services. A profile is a standard derived from a base standard by adding restrictions: it may require (or exclude) an element that is optional in the base standard; it may limit the valid entries under a heading; it may fix the cardinality of a repeating element; and so on. But, the profile cannot contradict the base standard; anything mandatory in the base standard remains mandatory in the profile. Thus, any product that complies with the profile will comply with the base standard. [1, 2]

One profile presented here is HDF-EOS, an EOS-specific adaptation of the very general Hierarchical Data Format. Of the many different file structures that are possible with HDF, HDF-EOS defines three (point, grid, and swath), each with spatial and temporal details alongside scientific data. In another example, FGDC's Metadata Content Standard has allowed several community-specific profiles to be defined, and in fact, the International Organization for Standardization's (ISO) Metadata standard was designed primarily for profiles. It defines several hundred elements of which fewer than 20 are required; the remaining elements are shared vocabulary (i.e., a dictionary) for building profiles. [3]

Related to profiles is the notion of *extensions*. These are elements added to the base standard by consensus among a certain community of users. As with profiles, extensions do not contradict the base standard – what's mandatory remains mandatory; products that fit the extended standard have everything needed by the base standard, and more. Nonetheless, by adding more loosely controlled, loosely defined elements to a standard, extensions may complicate the interoperability and maintenance of the standard.

For example, the earth imagery user community has extended the FGDC metadata content standard to more fully describe remotely sensed data by adding metadata elements such as the sensor model and the orbital platform, both of which the base standard doesn't provide [4].

4.3 Abstract vs. Implementation Standards

Standards and specifications for information systems are defined primarily at two different "levels of abstraction;" implementation specifications and abstract models [5].

- *Implementation specifications* tell software developers *how* to express information or requests within particular distributed computing environments (such as XML, Java, or the World Wide Web). Such standards define data formats, access protocols, object models, naming conventions, etc., in terms that are directly usable within the targeted computing environment.
    - *Implementation specifications* are the more immediately useful standards when they apply to one's chosen computing context. The data-format standards are implementation specifications, as are the eXtensible Markup Language (XML) encodings of FGDC, ISO, and other metadata standards seen in the Appendix, Section 3.0.

- *Abstract models* specify *what* information or requests are valid or required in principle, irrespective of individual computing environments. They define the essential concepts, vocabulary, and generic structure (type hierarchy) of computational services and information transfer. Although not directly usable to build data or software, these models set the stage for creating implementation specifications and for extending existing ones to new environments.
    - *Abstract models* provide well-known semantics that can support interoperability through translators or cross-reference tables. For instance, thanks to FGDC's Content standard, Z39.50's GEO profile can "normalize" any FGDC compliant metadata (regardless of actual record formats or field names) for external access – that is, map its internal data elements to the GEO field names for external access.
    - In general, consensus-based abstract models of data are often termed "content standards." They define the information elements and their intended meaning (semantics) independently of their syntax – that is, independent of how these elements may be encoded in files on disk or along a communications link. In principle, content standards allow different parties to communicate meaningfully by mapping their data element names to those of the content standard even when they use different formats for their data. This works well for fairly simple data structures such as the "parameter=value" pairs of many metadata files and catalog records. However, with more complex syntax or semantics, translating the abstract concepts of the content standard into the terms of a particular format often becomes an interpretation task requiring judgment calls, assumptions, and ambiguity. So in practice, content standards alone may not suffice for transferring complex data between different user communities without information loss or distortion.

4.4 Content and Format vs. Behavior and Interface

Table 4.4.1 shows that at each level of abstraction certain standards define the *interfaces* that allow different systems to work together or the expected *behavior* of software systems. This is the computation viewpoint, whose accent is on invoking services effectively and unambiguously. Other standards define the *content* of geospatial information or its *encoding* (or packaging) for accurate transfer between different processing systems. This is the information viewpoint, which emphasizes efficient, lossless communication [5].

Table 4.4.1 Viewpoints and Levels of Abstraction

	Service Invocation (computation viewpoint)	**Information Transfer (information viewpoint)**
Implementation specifications ("how")	Interface	Encoding (format)
Abstract models ("what")	Behavior	Content

For distributed computing, both of these viewpoints are crucial and intertwined. For instance, information content isn't useful without services to transmit and use it. Conversely, invoking a service effectively requires that its underlying information be available and its meaning clear. However, the two viewpoints are also separable: we may agree on how to represent information regardless of what services carry it; conversely, we may define how to invoke a service independently of how we package the information needed or conveyed by the service.

In a given context, either the computation view (implemented as interfaces) or the information view (implemented as formats) may take precedence. Tables 4.4.2 and 4.4.3 below show a few guidelines for prioritizing standards definition or adoption in certain contexts. In general, however, deciding which view to emphasize in a given setting is not straightforward.

Table 4.4.2 Criteria For Format Standards

Worry about a data format standard when …	Don't worry about a data format standard when …
communicate data with each other.	There's no reason for users of different formats ever to share information.
Each user group (or each user) uses a different format.	A user consensus already exists on one or a few non-proprietary data formats.
Available formats fail to convey all the information needed for proper use. (Thus users have to rely on implicit knowledge or ad-hoc notes to use the data.)	A practical, reasonably simple data format conveys all of the information users need.

Table 4.4.3 Criteria For Interface Standards

Worry about a service interface standard when …	Don't worry about a service interface standard (i.e. rely on FTP / FedEx) when …
Most users want the output of a few well-known processing operations, such as subsetting, filtering, transformations, etc.	Most users need direct access to raw data (as archived) for ad-hoc processing and analysis.
The intended applications are streamed or interactive – they only use parts of the available data at a given moment.	Most use of the data requires all of it (full size and detail) to be present simultaneously.
No one reasonably simple format will ever meet everyone's needs. (A service allows users to request the data they need in a format that fits it.)	Users have not begun to map their workflow to online database transactions or Web services.

Among the data standards reviewed in this report, GeoTIFF, Landsat Fast Format, and BUFR/GRiB are clearly file format standards; they specify an encoding and are silent on what access interface to use. HDF, HDF-EOS, and netCDF provide a software library to facilitate reading and writing data files, but they too are file format standards; they don't specify a format-neutral interface to a service. Table 4.4.4 compares the data models and software access libraries for a variety of data packaging standards.

Table 4.4.4 Data Models and Software Access Libraries

Data Format	***Logical Model***	***Physical Model***	***Software Access Libraries***
HDF	• Disk format, hierarchical, and similar to Unix file systems • Self-description provided in global and local (individual objects) attributesHeader describes disk structure with metadata & pointers • Usable for general scientific data storage; HFD4 data model contains: arrays, tables, raster image and text objects. HDF5 data model has HDF4-type objects imbedded within arrays and text attribute objects. • Will support extended (multiple machine) files	• XDR-based • Storage layout is contiguous (serial) or chunked (direct access) • Datasets consist of header attributes & data • Machine-independent	• C, C++, FORTRAN, Java
HDF-EOS	• HDF-based: Versions 4 and 5 • Provides standard for geolocation data map to science data . • Point Structure: model for sparce, randomly geolocated data • Swath Structure: model for data best organized by time, latitude or track parameter • Grid Structure: model for data organized	• (Same as HDF) • XDR-based • Storage layout is contiguous (serial) or chunked (direct access) • Datasets consist of header attributes & data • Machine-independent • Disk format is available to user	• C, C++, FORTRAN

Data Format	***Logical Model***	***Physical Model***	***Software Access Libraries***
	spatially and projected.		
netCDF	• Self-describing • Usable for general scientific data storage	• XDR-based • Storage layout direct access-- indexed • Datasets consist of header & data • Machine-independent • Disk format is hidden	• C, FORTRAN, Java, Perl, Python, Ruby. Tcl/Tk
GeoTIFF	• TIFF-based, with geolocation tags • Raster image data only • Multiple images can be stored in a single file. • Version 2 will support extended files	• Storage layout allows random access to pixels by band, strip, or tile	• C, Perl, Python, Java
	• Tailored to atmospheric data – point data • Based on sequential,, tape format	• Storage layout is serial • Dataset consists of header + data	• FORTRAN 77
	• Tailored to atmospheric data – gridded data • Based on sequential, tape format	• Storage layout appears to be serial – “messages” • Dataset consists of header + data	• Command-line translators to ASCII or IEEE binary
Format	• Multi-band image data	• Separate header and data files • Direct access to individual bands	• Users write their own software based on examples
	• Data model chosen by user. • Record, data types determined by specific platform.	• Different for every product • Machine dependent	• Custom software • Users must write their own

See Acronym List if needed

4.5 Web-based Data Service Standards

The World Wide Web is driving rapid development of format-neutral service interface standards. Examples particularly relevant to ESE data include the OpenGIS Web Coverage Service [6] and Web Map Service [7] and the Distributed Oceanographic Data System (DODS) [8].

The OpenGIS Consortium (OGC) Web Coverage Service (WCS) is likely to become an OGC specification in early 2003. It will provide access to images, imagery collections, and other systematic "fields" of values or measurements – usually arrayed on a 2D or 3D spatial grid. It fully describes the data's spatial location and its semantic content and allows clients to request subsets in space or along any of the data dimensions using a syntax based on either Uniform Resource Locators (URLs) or structured XML messages. The EOSDIS Core System (ECS) Synergy effort intends to provide WCS access to its large online data holdings ("data pools"); and the GLOBE educational project ("global learning and observations to benefit the environment:") has begun experimenting with WCS and WMS (next).

The OGC Web Map Service (WMS) provides access to rendered maps and pictures using a simple, spatial query syntax and common graphics formats (PNG, JPEG, etc.). Since its inception in early 2000, this interface has seen widespread implementation by many vendors, laboratories, and open-source efforts.

The Distributed Oceanographic Data Service provides format-neutral access to scientific datasets; its query syntax allows for "slicing" or "sampling" a dataset along any of its variable values. DODS originated at MIT and the University of Rhode Island (URI) in the mid-1990s; since then, it has seen a fair bit of implementation in the oceanographic community and among NASA DAACs. Recently, URI and NASA-DAACs have built "gateways" from DODS to WMS and WCS; and URI has begun defining two distinct successors to DODS: an "Open Source Project for a Network Data Access Protocol" (OPeNDAP) (tools for generic infrastructure protocols) and a "National Virtual Ocean Data System (NVODS)" (to supply oceanographic data and applications). [9]

(Notable Web-based services in the ESE environment include the University of Maryland's MOCHA project ("Middleware based on a code-shipping architecture") [10]; the Tropical Rainforest Information Center (TRFIC) at Michigan State University [11]; EOS-Webster at the University of New Hampshire [12]; and many others. However, these are not service interface standards but, rather, particular implementations of distributed systems. Although they provide a useful benefit to their users, they are not linked by a well-defined, published service interface standard; instead, they rely on tightly coupled components or on unpublished or proprietary interfaces.)

Finally, a number of vendors in the world of e-commerce have championed the notion of "Web Services" [13] consisting of the Web Services Description Language (WSDL) [14]; Simple Object Access Protocol (SOAP) [15]; and Universal Description, Discovery, and Integration (UDDI) [16]. These industry specifications have gained broad visibility and offer a lot of promise for Web-based data access; however, the dust is far from settling on this very active area of technology development.

Generally, the use of Web-based services is still only emerging in practical ESE work. The primary mechanism for information interchange in the ESE context remains the transfer of discrete files; it will take some time before Web-based services become a part of mainstream data access and distribution in ESE. Accordingly, this document treats format and content standards only for the near-term missions.

References:

[1] ISO TC211 (2001), "Geographic Information – Profiles" http://www.isotc211.org/protdoc/211n1134/211n1134.pdf

[2] Federal Geographic Data Committee (1998), Content Standard for Digital Geospatial Metadata: http://www.fgdc.gov/metadata/csdgm/ : Appendix D; Appendix E.

[3] Simon Cox (2001), Summary of some geospatial metadata standards: http://www.ned.dem.csiro.au/research/visualisation/metadata/geospatial/

[4] NASA (2001), Digital Earth Reference Model v0.5: http://www.digitalearth.gov/derm/v05/

[5] FGDC (2001), Content Standard for Digital Geospatial Metadata: Extensions for Remote Sensing Metadata: http://www.fgdc.gov/standards/status/csdgm_rs_ex.html

[6] Evans, John D. (2001), OGC Web Coverage Server (WCS), Discussion Paper #01-018: http://www.opengis.org/techno/discussions/01-018.pdf

[7] de La Beaujardière, Jeff (2001), OGC Web Map Server Interface Implementation Specification, version 1.1.1: http://www.opengis.org/techno/specs/01-068r3.pdf

[8] Distributed Oceanographic Data System (DODS): http://www.unidata.ucar.edu/packages/dods/

[9] Cornillon, Peter (2002), "DODS: OPeNDAP providing plug-and-play interoperability in a distributed data system," presentation at the 9th Assembly Meeting of the ESIP Federation, University of Maryland, College Park, May 15-17, 2002. http://www.esipfed.org/business/library/meetings/9th_fed_meeting/ppt/DODS.PPT

[10] The MOCHA project: Self-Extensible Middleware Architecture. http://www.cs.umd.edu/projects/mocha/

[11] Tropical Rain Forest Information Center (TRFIC). http://www.bsrsi.msu.edu/trfic/

[12] EOS-WEBSTER: Earth Science Data from the University of New Hampshire. http://eos-webster.sr.unh.edu/

[13] World Wide Web Consortium (W3C), 2002: Web Services home page. http://www.w3.org/2002/ws/

[14] World Wide Web Consortium (W3C), 2002: Web Services Description Language (WSDL) Version 1.2: W3C Working Draft 9 July 2002. http://www.w3.org/TR/2002/WD-wsdl12-20020709/

[15] World Wide Web Consortium (W3C), 2002: SOAP Version 1.2 Part 1: Messaging Framework. http://www.w3.org/TR/soap12-part1/.

[16] Universal Description, Discovery, and Integration (UDDI) of Business for the Web. http://www.uddi.org

5.0 Standards Evaluation

In order to objectively assess the data and metadata standards identified in Chapter 2 for the SEEDS near-term missions, an analysis is carried out to evaluate the standards according to many features or criteria. Furthermore, a user opinion interview/survey is conducted to gather user community's feedback on using the standards.

5.1 Evaluation Criteria

Many features or criteria can be used to evaluate the data and metadata standards identified. The intention of this study is not to identify one all-purpose standard but, rather, to identify appropriate use of the standards. For example, some standards are more suitable for transmission and archiving while others for analysis. For transmission and archiving, the most important features standards should have are semantic completeness, portability, self-description, extensibility, interoperability, etc [1]. For analysis, standards should have features such as ease of use, analysis tools support, etc. Many of these features and others are defined below.

1. Interoperability – Tools exist to translate to other standard formats with no information loss.

- Is there a defined relationship or semantic equivalence between this standard and other standards? *i.e.,* can the standard be broken into elements that have the same content as elements for other standards?
- Is the definition sufficiently precise to allow development of a translation algorithm between standards?
- What translation tools (well known) have been developed?

2. Availability – Source code for writing and reading data in the format is widely and publicly available.

- Is the source code for writing and reading data widely and publicly available?
- Is the software for reading and writing well documented?
- Are the search and order methods for data using the format well understood and established?

3. Portability – Data in this standard can be used on a variety of platforms or in a variety of applications (vendor support).

- Is the format sufficiently well defined so that data can be ported to new commonly used platforms with minimal effort?
- Is the format sufficiently well implemented that new applications can access the implementation with minimal effort?
- Can the standard be implemented on one platform and installed and tested on other platforms with minimal modification of source code? *i.e.*, machine dependent code is minimized.

4. Evolvability – A clear process for maintaining and evolving the standard exists.

- Is there a methodology for adding new features to the standard?
- Is there a software development process?
- Is there a standard for documentation?
- Is there an open process for evolution?

5. Extensibility – Support for extensions and profiles exists.

- Does the standard allow extensions or profiles to be developed?
- Are there extensions or profiles developed for the standard?

6. Self-describing - Files contain data descriptions along with the data.

- Can data in this format be read without a separate document detailing file contents?
- Can the data be described internally to facilitate development of applications?
- Does the format contain information to allow geospatial, temporal, and/or spectral subsetting?

7. Tools Support – Software tools are available to support the standard.

- Does the standard have freeware support?
- Does the standard have COTS (Commercial Off-The-Shelf) software support?

8. Completeness – The capacity to carry semantic descriptive elements of the data explicitly and unambiguously. Higher levels of completeness can reduce the user's dependency on outside information, implicit knowledge, or guesswork when interpreting and applying the data.

- Can the format carry everything users need to use the data correctly? *i.e.*, can the format convey the data's precise spatial location, its units of measure, the observation parameters (*e.g.,* spectral bands), accuracy estimates (error bars), and other elements needed to understand the data and apply it?

5.2 Data Standards Evaluation

Using the standards evaluation criteria defined above, Tables 5.2.1 through 5.2.8 analyze and compare data standards in use in heritage missions and other ESE missions.

Table 5.2.1 Data Standards Interoperability

Data Standard	Evaluation Questions		
	Is there a defined relationship or semantic equivalence between the standard and other standards?	**Can translation algorithms be developed easily?**	**What Translation Tools (well known) developed?**

HDF	Yes. Since HDF can contain general scientific data, it encompasses all the other standards.	Yes, HDF has a well-documented software API.	GIF <-> HDF5
			HDF4 <-> HDF5
			Ensight6 -> HDF5

HDF-EOS	Yes. As a superset of HDF, it also encompasses the other standards.	Yes, Point, Grid Swath add-on structures are well-documented.	GIF <-> HDF5
			HDF4 <-> HDF5
			Ensight6 -> HDF5

GeoTIFF	Yes, for image-based standards; no, for non-image standards.	Yes. Public domain API library partially documented.	Lots of converters for TIFF; also GeoTIFF tag read & write Specialized converters for L7, MODIS, MISR, ASTER
Fast Format	No	No. No API or library exists.	No
Native Binary	Depends on the standard. Most are specific to the application.	Depends on the standard, but usually not, unless specific efforts are made to document and publish an API.	No, You have to write your own translation tool
netCDF	Yes. Since netCDF can contain general scientific data, it encompasses all the other standards.	Yes. Net CDF has a well-documented API.	-> HDF -> Matlab5
BUFR/GRiB	Yes – translation of meteorological parameters to other formats is possible, with no loss of content. No for non-meteorological	Yes	BUFR -> CDF

Data Standard	Evaluation Questions		
	Is there a defined relationship or semantic equivalence between the standard and other standards?	**Can translation algorithms be developed easily?**	**What Translation Tools (well known) developed?**
	standards.		

See Acronym List if needed

Table 5.2.2 Data Standards Availability

Data Standard	Evaluation Questions		
	Source code for writing and reading data widely available?	**Read/write software well documented?**	**Format well described to facilitate application development?**
HDF	Yes	Yes	C, C++, Fortran, and Java interfaces exist. Applications must use one of these interfaces to access the data
HDF-EOS	Yes	Yes	C, C++, Fortran, and Java interfaces exist. Applications must use one of these interfaces to access the data
GeoTIFF	Open source libraries; many COTS and freeware applications available	User interface well documented	TIFF format well documented. COTS venders sometimes use variations of the standard.
Fast Format	No	No	No
Native Binary	Not always	Not always	Not always
netCDF	Yes (C, C++, FORTRAN, Perl)	Yes	Yes
BUFR/GRiB	There are few slightly different read and write software from different organizations or countries	Not always	Not always

See Acronym List if needed

Table 5.2.3 Data Standards Portability

Data Standard	Evaluation Questions		
	Portable among commonly used platforms?	**Format is sufficiently well implemented that new applications can access the implementation with minimal effort?**	**Standard can be implemented on one platform and installed and tested on other platforms with minimal modification of source code?**
HDF	Precompiled HDF libraries for a variety of popular platforms such as AIX, Cray HP,SGI,Sun, Linux and Windows.	Yes	Yes
HDF-EOS	Precompiled HDF-EOS libraries for a variety of popular platforms such as AIX, HP, SGI, Sun, and Linux.	Yes	Yes
GeoTIFF	Works on common OS's (Linux, Unix, Windows). Designed to be portable, but need some knowledge of specs.	Need some knowledge of the specs., Need understanding of geotags to develop applications.	Yes
Fast Format	Yes	No	Yes
Native Binary	Usually not	No	No
netCDF	All major OS’s: Winx, Unix, Linux, MacOS	Yes	Yes
BUFR/GRiB	YES	A generalized application would require in depth knowledge of all variants, which is not easy to obtain	YES

See Acronym List if needed

Table 5.2.4 Data Standards Evolvability

Data Standard	Evaluation Questions			
	Is there a methodology for adding new features?	**Is there a software development process?**	**Is there a standard for documentation?**	**Is there an open process for evolution?**
HDF	NCSA is a currently active and outside funded group whose purpose is devoted to the HDF project. They manage development schedules and are open to suggestions from users. They are funded from a variety of sources.	Yes, HDF library is funded and developing software.	Yes, HDF library follows an internally defined standard for their documentation.	Yes, HDF group allow input from outside users
HDF-EOS	Support is a contract from NASA. They respond to suggestions from users. It is NASA's decision on how long to support the contract and whether to supply money for development as well as maintenance.	Yes, HDF-EOS library is funded and developing software.	Yes, HDF-EOS library follows an internally defined standard for their documentation.	Yes, HDF-EOS group allow input from outside users
GeoTIFF	Maintained by JPL..No formal process, i.e. Standards committee. The	Yes	Yes	OpenGIS, but no formal process. Work on the GeoTIFF v2.0 spec has been

Data Standard	**Evaluation Questions**			
	Is there a methodology for adding new features?	**Is there a software development process?**	**Is there a standard for documentation?**	**Is there an open process for evolution?**
	standard can be modified by others.			slow recently, with some recent efforts
Fast Format	No	No	No	No
Native Binary	No	No	No	No
netCDF	Yes, through Unidata	Yes	Yes	Yes, informally through Unidata
BUFR/GRiB	YES	NO	Appears so, WMO issues Tech. Docs. on these formats	The WMO CBS approves changes to the format and maintains a software registry

See Acronym List if needed

Table 5.2.5 Data Standards Extensibility

Data Standard	**Evaluation Questions**	
	Does the standard allow extensions or profiles to be developed?	**Are there extensions or profiles developed for the standard?**
HDF	Yes	HDF-EOS is a profile which was developed.
HDF-EOS	This is a profile of HDF	No
GeoTIFF	Yes. New projections can be added. Multiple-band GeoTIFFs allowed. GeoTIFF 2.0 will allow external files.	None that are not part of unofficial list of projections
Fast Format	No	No
Native Binary	No	NO
netCDF	Yes	Yes, e.g., MINC: (Medical Image netCDF)
BUFR/GRiB	YES	Not sure

See Acronym List if needed

Table 5.2.6 Data Standards Self-Describing

Data Standard	**Evaluation Questions**		
	Is data able to be stored so that it can be read without a separate document detailing file contents? .	**Can the data be described internally to facilitate development of applications?**	**Does the format contain information to allow subsetting?**
HDF	Data can be stored so that it is self-describing. There are no restrictions in the standard though to prevent developers from using names such as Variable1.	Data can be described with enough detail to allow applications to process data appropriately. For instance, scale factors may be included but it is developer dependent on how to do this. As a result, generic applications are limited in their scope. Applications developed for a specific data set can be very precise.	Yes, information can be supplied to allow subsetting, but there is not a requirement to do so in a consistent way. Subsetting by selecting selected data fields can easily be done on any HDF file.
HDF-EOS	Data can be stored so that it is self-describing. There are no restrictions in the standard though to prevent developers from using names such as Variable1.	Data can be described with enough detail to allow applications to process data appropriately. For instance, scale factors may be included but it is developer dependent on how to do this. As a result, generic applications are limited in their scope. Applications developed for a specific data set can be very precise.	Because of the profile, subsetting along certain geolocation fields can be done. Individual developers can break this process by not following the profile (there is no internal checking done).
GeoTIFF	Geotags and image specs.are in an ASCII header. Need library to access contents.	Information can be extracted at a pixel level. Geolocation and image info. is available	Yes

Data Standard	Evaluation Questions		
	Is data able to be stored so that it can be read without a separate document detailing file contents? .	**Can the data be described internally to facilitate development of applications?**	**Does the format contain information to allow subsetting?**
		through the interface.	
Fast Format	No	No	No
Native Binary	No	No	No
netCDF	Yes	Yes, it was designed to be self-describing. CDL (schema-like) files are used to create files initially, but are not needed thereafter.	Not inherently. However, community-defined netCDF conventions can be used to write code that allows subsetting
BUFR/GRiB	No, you need tables to interpret the data	YES, but codes are used to describe projections, geophysical parameters, etc., so need to know these codes to interpret the data	YES, but need the tables

See Acronym List if needed

Table 5.2.7 Data Standards Tools Support

Data Standard	Evaluation Questions	
	Does the standard have freeware support?	**Does the standard have COTS support?**
HDF	Yes, NCSA tools ImageMagick HDFLook HDFExplorer WIM H5View	Yes, PCI ENVI ER Mapper ERDAS-Imagine HDF Explorer IDL ImageMagick MATLAB Mathematica NCL Noesys PV-Wave WIM
HDF-EOS	Yes, EOSView HE5View Webwinds	Slow: ENVI IDL MATLAB Noesys
GeoTIFF	Open-source GIS tools (GRASS) and libraries for C (libgeotiff), Java (JAI), Python, etc.	Widespread: PCI-Geomatica RSI-ENVI ESRI-ArcView SoftDesk-AutoCAD ER Mapper ERDAS-Imagine Laser-Scan MapInfo MicroImages Intergraph-GeoMedia ENVI/IDL[1]
Fast Format	GRASS, GDAL, OSSIM (Simple format, so manual import is common)	Moderate: PCI ENVI ER Mapper ERDAS-Imagine MicroImages

[1] Any tool that reads a TIFF file should also read a GeoTIFF file (though most will complain about the extra "unsupported" tags). The packages listed here are those that use the additional information contained in a GeoTIFF file to geolocate the data.

Data Standard	**Evaluation Questions**	
	Does the standard have freeware support?	**Does the standard have COTS support?**
		(Simple format, so manual import is common)
Native Binary	No	No
netCDF	DODS GMT Linkwinds GrADS VisAD etc.	AVS Environmental workbench IDL interface IRIS Explorer Module MATLAB NCAR graphics Noesys PPLUS PV-Wave Silver Dicer WXP
BUFR/GRiB	Bufkit, GrADS	Limited

See Acronym List if needed

Table 5.2.8 Semantic Completeness

Data Standard	**Completeness Questions**	
	Can the format convey the data's precise spatial location?	**Can the format convey the units of measure, the observation parameters (e.g. spectral bands), accuracy estimates (error bars), and other elements needed to understand the data?**
HDF	Yes	Yes
HDF-EOS	Yes	Yes
GeoTIFF	Yes (basically)	No
Fast Format	Yes (basically)	No
Native Binary	N/A	N/A
netCDF	Yes	Yes
BUFR/GRiB	Yes with qualifications	Yes with qualifications

See Acronym List if needed

Based on the analysis of data standards using the eight criteria defined and the evaluation questions, each standard was giving a rating (low, medium and high) for each criterion. This rating is based on the answers to the evaluation questions. If a data standard can satisfy all the evaluation questions, a high rating is giving. If a data standard cannot satisfy any of the evaluation questions, a low rating is giving. If a data standard can satisfy some of the evaluation questions, a medium rating is giving. Table 5.2.9 summarizes the results.

Table 5.2.9 Data Standards Evaluation

Criteria	Data Packaging Standards						
	HDF	HDF-EOS	GeoTIFF	netCDF	Fast Format	BUFR/ GRiB	Native Binary
Interoperability	High	High	High	High	High	Medium	Medium
Availability	High	High	High	High	Low	Low	Medium
Portability	High	High	High	High	High	Medium	Medium
Evolvability	High	High	Medium	High	Low	Medium	Low
Extensi-bility	High	Medium	High	High	Low	Medium	Low
Self-describing	High	High	Medium	High	Low	Low	Low
Tools Support	High	High	High	High	Medium	Medium	Low
Completeness	High	High	Low	High	Low	Medium	N/A

Table 5.2.9 shows that:

- HDF and netCDF receive high ratings for all eight criteria. This suggests that HDF and netCDF are good candidates as transmission or archive standards.
- Many standards, including HDF, HDF-EOS, GeoTIFF, and netCDF receive high ratings for Tools Support. This suggests that these standards can be used as analysis standards. Different user communities may prefer one standard over the others based on their familiarity with the standard and the simplicity and ease of use of the standards.

Metadata and Documentation Standards Evaluation

This section analyzes five metadata standards and one documentation standard used in heritage missions and related user communities:

- The Federal Geographic Data Committee's Content Standard for Digital Geospatial Metadata;
- ISO's draft standard on geographic metadata (ISO 19115);
- The EOSDIS (Earth Observing System Data and Information System) Core System Core Metadata Standard;
- The Directory Interchange Format (DIF) of the Global Change Master Directory (GCMD);
- Global Information Locator Service (GILS) and Dublin Core elements;
- The EOSDIS Data Gateway (or Information Management System (IMS) V0) Guide.

The above metadata and documentation standards are analyzed according to the same criteria used for data format standards. Table 5.3.1 shows the analysis results.

Table 5.3.1 Metadata and Documentation Standards Evaluation

Criteria	Metadata Format					
	ISO 19115	FGDC Content Standard		GCMD	GILS	Guide
	High: FGDC CSDGM, GCMD-DIF, etc. to be redefined as profiles of ISO-19115.	High: GILS, ECS, CIP (via GEO profile of Z39.50 API). To be reconciled w/ ISO 19115.	High: GCMD, FGDC compatible	High: ->ISO ->FGDC ->Dublin Core ->ANZLIC	High: FGDC, ECS, CIP (via Z39.50 API).	Medium: ->FGDC
Availability	Low: draft on private ISO Website; final std. will cost	High: standard is on FGDC website	High: SDP and MDT Toolkits on HDF-EOS website	DIFs & DIF authoring tool available on web	High	High: Guide template available on web
	High content w/ xml schema	High: content w/ xml schema	High: Portable, can be created w/ text editor	High: can be created w/ text editor	High	be created w/ text editor
Extensibility	High (Built for profiles & extensions)	Moderate (extensions exist)	High: Can be extended	be extended	N/A (not applicable)	High: Can be extended

Criteria	Metadata Format					
	ISO 19115	FGDC Content Standard		GCMD	GILS	Guide
	High (many orgs. maintain it)	Moderate	Medium: Will be supported by ESDIS/ EMD	High: GCMD group maintain the DIFs	Moderate	sci/ops maintain the Guide
Self-describing	Moderate	Moderate	Medium: Some attributes require documenta-tion to understand	High: Collection level metadata self-describing, no granule metadata	Moderate	Guide template is self-describing
Tools Support	Low: Emerging translators to/from FGDC; etc. No vendor support yet	High: fgdcmeta + xtme SMMS Metamanager, MetaStar mp, cns Many vendor support	High: SDP toolkit Metadata works QAMUT EDG TerraWhom ECHO	High: Authoring tool DIF to XML Open access API Science keywords interface	High: Many: Isite, Meta Manager, Meta Star most common No vendor support	Medium: Converter Card creator Test bed Verifier No vendor support

Criteria	Metadata Format					
	ISO 19115	FGDC Content Standard		GCMD	GILS	Guide
			UFM No vendor support	support		
Completeness	High: carry semantic descriptive elements of the metadata	High: carry semantic descriptive elements of the metadata	semantic descriptive elements of the metadata	Low: Only semantics descriptive elements for collection metadata	High: carry semantic descriptive elements of the metadata	N/A (not applicable)

See Acronym List if needed

As shown in Table 5.3.1, all five metadata standards and the documentation standard receive similar ratings for most of the criteria used. 1This is because the metadata standards analyzed, FGDC CSDSM, ISO 19115, and ECS data model, are all based on each other. As described in the Appendix, Section 3.0, ISO 19115 was originally based on FGDC CSDSM version 1.0 and current FGDC "is consistent with the emerging ISO draft standards". FGDC will adopt ISO 19115 standard when it becomes final in 2002. The ECS data model is FGDC compliant, and FGDC Remote Sensing Extensions adopted many of the attributes in the ECS data model. Since ISO 19115 is an international metadata standard, it seems natural for FGDC and ECS data model to adopt ISO 19115 when it becomes final. ISO 19115 receives a low rating for "Availability" and "Tools Support" because the final ISO19115 standard has not been published.

GCMD is a metadata (collection level only) standard for on-line catalog access. Therefore, it is different from the ISO19115, FGDC CSDGM, and ECS data model. GCMD has been widely used in NASA Earth Sciences and GCMD DIF has been cross-mapped to ISO 19115 and FGDC CSDSM. GCMD receives a low rating for "Completeness" because it has only semantic descriptive elements for collection metadata.

Guide document standard and GIL are interoperable with the FGDC metadata clearinghouse via Z39.50 API. However, Guide document standard is more suitable for Earth Science data sets.

5.4 User Surveys

In conjunction with the analyses of standards described above, we also conducted a user opinion survey (see the Appendix, Section 4.0) on the data and metadata standards used in heritage and other missions to gain feedback from the user community. Not all of the criteria used in the standards analysis above were used in the survey because the survey was conducted before the criteria selection was refined. Also, there are several criteria used in the survey but not in the analysis. These criteria are more subjective than the more refined criteria used in the standards analysis and are defined below.

- *Ease of use* for producers.
- *Ease of use* for consumers.
- *Acceptability* - Format is acceptable to a broad cross-section of likely users of the products.
- *Suitability* - Has the proper descriptive power or precision for the task.
- *Survivability* - The ability to be used by the community for many years.

There were a total of 45 survey respondents. Twenty surveys were returned from attendees of the NASA Science Data Processing Workshop, February 2002. Twenty-five surveys were collected from EOS User Working Group members at GSFC, LaRC, JPL, EDC, NSIDC, and ORNL DAACs (See Acronym List) and from other users. The survey results and some relevant statistics are summarized below.

5.4.1 Data Format Standards

Respondents were asked the question: "What weight should NASA give to the following criteria in evaluating a standard?" They were then asked to rate defined criteria (exact definitions are shown in the Questionnaire found in the Appendix, Section 4.0) with respect to "what weight NASA should apply when evaluating a standard," using a scale from one to six, where one was the lowest and six was the highest. Users gave "Ease of Use for Consumers" the highest rating (5.6) while "Evolvability" and "Ease of Use for Producers" received the lowest ratings. The statistics are summarized in Table 5.4.1.

Table 5.4.1 Survey Ratings of Attribute Importance

Statistic	Criteria								
	Ease of Use For Producer	Ease of Use For Consumer	Survivability	Acceptability	Availability	Portability	Evolvability	Suitability	Interoperability
Average	4.3		5.3	5.1	5.1	5.4	4.6	4.6	4.8
Mode	4	6			6	6	4	5	
Sample Size	41		42		41	42	41	39	41

Survey respondents were then asked to rate different data format standards using the set of criteria from the previous question. They were also ranked on a scale from one to six, where one was the lowest and six was the highest.

The survey of data standards used in the heritage missions indicated that users are most familiar with the HDF data standard. Thirty-five of the 45 total respondents were familiar with and rated HDF, while only one respondent rated Fast Format and only three rated BUFR. The sample sizes of the responses concerning Fast Format and BUFR were so small that these two formats were deleted from the results. Overall, the number of respondents is small (total respondents is 45, however, the sample size for a particular standard and particular criteria ranges between 5 and 35 as shown in Table 5.4.2); thus, drawing decisive conclusions from the survey is difficult. For example, approximately one-third of the people surveyed were not familiar with more than three data standards, and in many cases they gave the standard they were most familiar with the highest rating. However, the survey still shows some interesting findings.

- Binary format received the highest rating for most of the criteria used including Interoperability, Acceptability, Availability, Survivability, Evolvability, and Ease of Use. However, Binary was rated the lowest for Portability.
- HDF and netCDF were rated highest for Suitability and Portability. However, they were rated lowest for Ease of Use. This is due to the steep learning curve as indicated by the people surveyed.
- GeoTIFF was rated the second highest for Ease of Use. It was also rated high for Interoperability and Acceptability. However, it was rated the lowest for Suitability, Availability and Evolvability.

A summary of the statistics is shown in Table 5.4.2.

Table 5.4.2 Data Standards Survey Evaluation

	Type of Statistic	Data Format				
		HDF	HDF-EOS	netCDF	GeoTIFF	Binary
Interoperability	Average			3.6	4.5	4.9
	Mode	6	4		5	
	Sample Size	28	24	10	8	
Acceptability		4.4		4.1	4.3	4.8
	Mode	6	3	4, 6	6	5, 6
	Sample Size	34			9	13
	Average	4.7	4.2		3.4	
	Mode	6	6	6	3, 4	5
	Sample Size	35	15	15	9	12
Portability	Average	4.8	4.5	5.1	4.7	4.6
			6	6	6	5
	Sample Size	34			11	12
Evolvability	Average	4.4	4.1	3.8		
	Mode	6	3, 6	6	4	6
	Sample Size	20	20	5	6	9
Suitability			5.1	4.6	3.9	4.1
	Mode	5			4	5
	Sample Size	28	25	9		
Ease of use for Consumer	Average	3.5	3.6	3.6	4.8	5.0
	Mode	4	3	2	4, 6	5
	Size		27	12	10	13
	Average	4.7		4.7	4.7	5.5
	Mode	5	4, 5	5, 6	4	6
	Size	29	24	12	10	13

Table 5.4.3 summarizes the results of four essay type questions from the opinion survey. An interesting result is that the HDF format received the most responses for all of the questions. About 60% of people surveyed listed that they had success with HDF and about 30% of the people listed HDF as a standard they foresee being used in the future. However, about 25% of the people surveyed listed that HDF is an impediment to their work. HDF-EOS, netCDF, and binary also received high ratings as a standard with which respondents had success (about 43% had success with HDF-EOS, 27% had success with Binary, and 23% had success with netCDF).

A detailed analysis of the results indicates that most of the respondents who had success with HDF are data producers from the computer/data manipulation areas (75% of them had success), atmospheric scientists (69%), and oceanographers (50%).

Respondents who had problems with HDF are atmospheric scientists (25% of them had problems), computer/data manipulation specialists (25%), and environmental scientists (20%). The most common complaint was that HDF was too complicated, making it difficult to learn (7 out of the 12 respondents), and that there was a lack of available tools (5 out of the 12 respondents). One atmospheric scientist seemed to capture this idea fairly well: “HDF, HDF-EOS, and netCDF are all initially more difficult, because they are more complex. But it isn't really an impediment as long as you have the right tools.” Other impediments included huge file size, slow conversion from HDF4 to HDF5, and performing compression in HDF libraries. Similar complaints were listed for HDF-EOS, but also lack of support by RSI and cryptic failure messages were mentioned.

For other data formats, one atmospheric scientist said that GeoTIFF showed a dependency on machines. Another atmospheric scientist called BUFR, “ancient and primitive.” The lone comment from an oceanographer for the GRiB format labeled it as difficult to use (worse than HDF). And one data producer said that binary was platform dependent and hard to verify.

Respondents who recommended HDF as a future standard are data producers (33% of them recommended), atmospheric scientists (25%), and 17% of oceanographers. Of the 21 respondents that commented on formats they foresee emerging in the future, eight (four atmospheric scientist, three data producers, and one environmental scientist) foresaw some form of HDF (HDF4, HDF5, or HDF-EOS) for the future because it is a powerful and versatile format. Four respondents (two data producers and two atmospheric scientists) felt some form of HDF was inevitable/mandated for the future.

Concerning other data formats, one oceanographer foresees and recommends GeoTIFF because it’s easy to use, one atmospheric scientist says that ASCII is obvious for small data sets, and two respondents (an atmospheric scientist and an oceanographer) foresee/recommend netCDF because IDL supports it or because younger scientists are adopting it so it’s growing in popularity. An oceanographer and an atmospheric scientist said that they foresee binary in the future because it has many tools and is widely available and because it is easy to use. Two scientists (atmospheric and environmental) recommend binary because it is simple, while one data producer recommends binary because many models require binary inputs.

Other comments on what the respondents foresee/recommend include: 1) atmospheric scientist – “Simple storage layer coupled with sophisticated connectivity layer. Let data producers have more freedom to use most appropriate format but force connectivity rules," 2) environmental scientist – "Something more easily accessible across platforms, software packages," 3) data producer – "Provide free, portable tools for ingesting and reformatting HDF and HDF-EOS. End-users will want to easily convert to formats such as ERDAS, ENVI/IDL, PCI, and Arcgrids."

Table 5.4.3 Summary of Survey Essay Questions

Question	Data Format				
	HDF	**HDF-EOS**	**netCDF**	**GeoTIFF**	**Binary**
Data Format Used Successfully	**27**	**19**		**2**	**12**
Data Format was Impediment	**12**	**9**	**1**	**1**	
Foresee Emerging	**13**	**8**	**4**	**1**	**4**
Recommend	**13**				**3**

5.4.2 Metadata Format Standards

For the metadata section, respondents of the user questionnaire were given a series of questions identical to those asked in the data format section of the survey. Of the 45 total surveys collected, only 16 people responded to the metadata questions. Thirteen of those were returned from attendees of the NASA Science Data Processing Workshop, February 2002. Three of those surveys were from EOS User Working Group members.

As in the data format portion of the survey, respondents were given a series of eight questions with a set of defined criteria as applied to particular metadata formats. They were asked to rate each metadata format (see headings in Table 5.4.4) with respect to the criteria using a scale from one to six, where one was the lowest and six was the highest.

The survey of metadata standards used in the heritage missions indicates that users are most familiar with the ECS Data Model, with as many as 11 responses out of the 16 total metadata responses, while only one respondent rated ISO and only two rated GCMD. The sample sizes of the responses concerning ISO and GCMD were so small that these two metadata formats were deleted from the results. Overall, the number of respondents is extremely small (as mentioned above) with the sample size for a particular standard and particular criteria ranging from 3 to 11 as shown in Table 5.4.4; thus, drawing decisive conclusions form the survey is difficult.

From the limited samples shown in Table 5.4.4, the FGDC content metadata and ECS data model received comparable ratings for most of the criteria. The FGDC content metadata received a little higher ratings than the ECS data model on Acceptability, Availability, Evolvability, and Survivability, while the ECS data model received a little higher ratings than the FGDC metadata model on Portability and Interoperability.

Table 5.4.4 Metadata Standards Survey Evaluation

	Statistics	**Metadata Format**	
		FGDC	**ECS**
Interoperability		**3.0**	**3.4**
	Sample Size	**3**	
Acceptability	Average	**4.0**	**3.8**
	Sample Size	**5**	**10**
	Average	**4.0**	
	Sample Size	**4**	**11**
	Average	**3.0**	
	Sample Size	**4**	**10**
	Average	**5.0**	
	Sample Size	**2**	**9**
	Average	**4.4**	
	Sample Size	**5**	**11**
Consumer	Average	**3.4**	
	Sample Size	**5**	**10**
	Average	**5.6**	
	Sample Size	**5**	**9**

Table 5.4.5 summarizes the results of four essay type questions from the opinion survey. Sample sizes are very small as only 11 respondents answered the first question, 4 respondents answered the second and third questions, and 3 people answered the fourth question. The most significant observation is that 8 (73%) of the 11 respondents listed that they had success with the ECS Data Model. Only 27% claimed success using the FGDC Content Standard. 27% of the 11 respondents said that the ECS Data Model was an impediment, while 9% claimed the FGDC content standard was an impediment to their research. A detailed analysis of the results indicates that the respondents who

claimed success with the ECS data model are data producers (5) and oceanographers (2). Some of the data producers (3) who claimed success with the ECS data model also indicated that they had problems with the ECS data model. Most of the complaints related to the ECS data model are that it is too complex, not flexible, not consistently applied, and that there is not enough tool support. Respondents who claimed success with the FGDC content standard include one data producer, one environmental scientist, and one atmospheric scientist. The environmental scientist who had used FGDC successfully also indicated that FGDC is an impediment in that there are too few tools and little portability, documentation, and interoperability. In terms of metadata that the respondents foresee in the future, one suggests adopting ISO 19115 to replace the current FGDC and adopting the FGDC extensions for remote sensing based on the ECS data model. One respondent recommends XML standard descriptions and defining XML DTD/schema for all the specific applications. One respondent suggests refining the ECS data model, dropping most of the groups/classes, and attaching metadata to files.

Table 5.4.5 Summary of Metadata Survey Essay Questions

Question	Metadata Format Was Listed	
	FGDC	ECS
Data Format Used Successfully	3	7
Data Format was Impediment	1	3
Data Format Foresee Emerging	1	2
Data Format Recommentd	1	

References:

[1] Jim Frew, 1998, How to Think about Data Formats, http://spso.gsfc.nasa.gov/diss/Meetings/19981005/frew_pres_9810.html.

6.0 Summary

We have surveyed the standards for data and metadata that are in use in heritage missions or under consideration by the missions expected to be in formulation in the near-term. Lessons learned from heritage missions, some of the NOAA missions, and STDS have been reviewed. In particular, data and metadata standards in use in heritage missions and other EOS missions have been analyzed using a suite of criteria. Simple statistics and results from our user interview/survey to gather data producers' and data users' feedback on data and metadata standards are presented.

The highlights of lessons learned from heritage missions and other standards are summarized below.

- Multiple data distribution formats are used for some heritage missions, such as Landsat-7 and QuikSCAT/SeaWinds, to satisfy the diverse requirements from the user communities. Many NOAA POES missions use multiple data distributions formats to give users the flexibility to select the best data formats for their applications. This adds to the workload of NOAA agencies, such as NOAA NESDIS and NOAA NCDC, requiring them to develop and maintain different data translation tools in order to support different requirements from their user communities.
- Many different versions of HDF-EOS have been implemented for Terra data products, thus, creating problems for data interchange between mission instrument teams and users because different readers may be needed to read different implementations of the HDF-EOS data format. New EOS missions have realized this problem. For example, the Aura mission instrument teams have decided to adopt a uniform set of HDF-EOS file format guidelines so that data products from any Aura instrument are easily interchanged, *i.e.*, the same set of tools and I/O routines can be used for all of the Aura data products.
- An important lesson learned from several missions, including Jason-1, SeaWiFS, SeaWinds, and ACRIM, is to provide good user support and experienced help desk for HDF-EOS implementation and usage. Many missions indicated that the "handholding" should not be underestimated.
- The Spatial Data Transfer Standard (SDTS), the national spatial data transfer mechanism for all U.S. Federal agencies, fell short of its ambitious goals and the marketplace was slow to accept and support it. The SDTS experience illustrates the importance of keeping pace with technology and market trends and emerging expectations, even after capturing initial requirements.

We devised eight standards criteria in order to objectively evaluate data and metadata standards. The results from the analysis of data format standards and metadata standards using the eight evaluation criteria are summarized below.

- HDF, and netCDF received high ratings for the evaluation criteria such as Interoperability, Availability, Evolvability, Portability, Extensibility, Tools

Support, Completeness, and Self-describing. Many standards, including HDF, HDF-EOS, GeoTIFF, and netCDF received high ratings for Tools Support.

- BUFR/GriB and Fast Format in general were rated low to medium for many evaluation criteria, mainly Self-describing, Availability, and Completeness.
- Native binary received low ratings for Evolvability, Extensibility, Self-describing, and Tools Support (Chapter 5). Based on our analysis, native binary does not constitute a data standard and it cannot be compared alongside open consensus standards.
- The metadata and documentation standards analyzed received similar high to medium ratings for most of the evaluation criteria as many of the metadata standards, such as FGDC CSDGM, ISO 19115, and ECS data model, are all based on each other. We note that metadata standards are converging on the ISO 19115 when it becomes final in the near future.

We conducted a total of 45 interviews and surveys of data users from the EOS User Working Group members at different DAACs and of data producers/users from the 2002 NASA Science Data Processing Workshop. Although the sample size is not large, the interview/survey results illustrate feedback from data users and data producers on the data and metadata standards currently in use in NASA missions. Statistics and results are described in Chapter 5. A summary of the statistics/results is presented here.

- All of the users/producers answered questions related to data format standards. Only one-quarter of the users/producers answered questions related to metadata standards. This indicates that data producers and data users care more about (or are more familiar with or is more relevant to them) data format standards than metadata, and that many of them also have strong feelings about the data format standards.
- Users/producers are most familiar with the HDF data standard as 35 of the 45 total respondents were familiar with, and rated HDF, while only one respondent rated Fast Format and only three respondents rated BUFR. Many of those interviewed/surveyed are not familiar with multiple data formats, with only one-half of the respondents familiar with more than two data standards. In many cases they gave the standard they were most familiar with the highest rating. The results, therefore, may be biased.
- The interview/survey results show that HDF and netCDF were rated highest for Portability and Suitability and lowest for Ease of Use. However, respondents did not give high ratings to HDF and netCDF on Interoperability, Acceptability, Availability, and Evolvability. On the contrary, respondents rated binary format the highest for Interoperability, Acceptability, Availability, and Evolvability. This could possibly be because the producers'/users' understanding of these criteria are different from what we described in Chapter 5. This could also be because users are more familiar with Binary format and favor Binary format rather than the HDF or netCDF formats.
- The majority (60%) of respondents indicated that they had success with HDF and about one-third of the respondents recommend HDF as a future standard for

NASA mainly because it is a powerful and versatile format. However, about one-quarter of the respondents also point out that HDF was an impediment to their work because HDF was too complicated, making it difficult to learn, and that there was a lack of available tools.

- For the metadata standards surveyed, respondents are most familiar with the ECS Data Model, with as many as 11 responses out of the 16 total metadata responses, while only one respondent rated ISO and only two respondents rated GCMD DIF. The ECS data model and the FGDC content metadata received comparable ratings for most of the criteria. This result is correspondent to the results derived from the standards analysis.
- For future metadata standards, some respondents recommend adoption of the ISO 19115 to replace the current FGDC and adoption of the FGDC extensions for remote sensing based on the ECS data model. Others recommend XML standard descriptions for metadata and refining the ECS data model.

7.0 Conclusions

Recommendations on data interface standards, data packaging standards, metadata standards, documentation standards, and associated activities for the near-term missions are summarized below. Acknowledging that there are several levels of requirements/guidelines, the following keywords are used to differentiate between them.

- *must* - This is mandatory.
- *should* - This guideline is mandatory *except* where valid reasons exist to allow for its modification. Care should be taken in modifying or ignoring these guidelines.
- *may* - This is a guideline which, while the NTMS group suggests it is worthwhile, is not mandatory.

7.1 Data Interface Standards Recommendations

1. For interface standards, data services based standards will become increasingly important. For ESE data, the leading definitions for such data delivery and interchange services are the OpenGIS Consortium (OGC) and the Distributed Oceanographic Data System (DODS). However, for the near-term missions, the preferred mode of delivering data remains the transfer of discrete files. In such case, the file format itself is critical to the interchange standard.
2. Web Service standards and XML will have an impact on data, metadata, and interface standards in the future. SEEDS *should* direct/track developments in the science and business communities.

7.2 Data Packaging Standards

The SEEDS Near-term Missions Study (NTMS) group recommends that data be packaged in an interchange format for storage and be available in multiple data distribution formats. The interchange format is for sharing data among the ESE data systems components (*i.e.*, data centers including PI-managed Mission Data Centers, "Backbone" Processing Centers, Science Data Centers, Application Data Centers, Multimission Data Centers, and similar centers or systems). Data distribution formats are for end-users.

NTMS recommends the following for packaging of near-term mission standard products.

7.2.1 Data Distribution Formats Recommendations

1. Data distribution facilities *must* enable packaging of standard data products in multiple distribution formats.
2. Distribution formats *must* emphasize end-user needs and convenience.

Rationale

At present, and for the near future, most applications and end-user practices are file-based. In such use, the format of the data files or the API used to read the data files are key to data access. Some communities have coalesced around particular file formats or access tools. For the greatest success in reaching multiple application and science disciplinary uses, ESE must support the preferences of these communities. These distribution-packaging choices (most simply understood as the formats in which data are sent to users) allow users to have access to data in one of several well-used formats. NTMS finds that several missions (including Landsat-7, QuikSCAT/SeaWinds, and many NOAA missions) have successfully employed multiple data distribution formats to satisfy the diverse requirements of their user communities (see Chapter 3).

All "Backbone" Processing Centers, Science Data Centers, and Multimission Data Centers *must* support a limited set of distribution packaging standards based on their end-users' requirements. The choice of distribution packaging *must* be made with the target community in mind and governed by applicability to task and the convenience of end-users.

NTMS does not have particular recommendations for distribution packaging standards. Based on our study, GeoTIFF format, WMO BUFR and GRID formats, Landsat Fast Format, and the API standards of netCDF and HDF/HDF-EOS and others are appropriate distribution packaging options.

7.2.2 Data Interchange Formats Recommendations

1. Interchange data sets *must* use a recognized packaging standard. The choice of standard *must* emphasize completeness and self-description.

2. Most of the near-term missions have indicated an interest in using HDF/HDF-EOS or netCDF. We agree that HDF/HDF-EOS or netCDF are appropriate choices as interchange data formats among ESE data system components (*i.e.*, data centers).

3. Each appropriate ESE Near-term mission community *must* develop a profile of HDF/HDF-EOS or netCDF appropriate not only to the narrow needs of a particular mission but also to the wider needs of the allied community (See Chapter 4 for a discussion of profiles).

4. The development of each community's profile *must* be a process involving mission science teams, interested end-users, and experienced consultants. SEEDS NTMS has found that "community based" standards are more closely followed than standards imposed by outside forces.

5. Each community's interchange format profile *must* be as specific as possible to eliminate differences between data products and allow for the generation of simple data packaging tools.

6. If the HDF-EOS geolocation makes sense for a particular community, it *must* develop its interchange format profile based on this standard.

7. Each community *should* review other Near-term mission's community interchange format profiles and incorporate sections of overlap in their profile.

8. The interchange formats *may* be used as distribution formats.

Rationale

An interchange packaging standard among PI-managed mission-data centers and other ESE data systems components ensures that data are completely and correctly transferred. Use of standards for this interchange increases the flexibility of the ESE data systems. New components can join with the ESE data systems to provide data services without negotiating one-to-one interface agreements with each potential provider. The effect of using standard packaging methods will result in decreasing the complexity of the ESE data systems as a whole while increasing potential for participation and novel use of NASA Earth systems science data sets. A community-involved process for approving or developing these data packaging standards ensures that the standards are appropriate and reliable. The choice of interchange packaging standards must consider completeness and correctness of representing data and emphasize the self-descriptiveness and long-term stability of the standards.

While HDF/HDF-EOS and netCDF, with a community-developed profile, are not the only possible candidates for an interchange standard, at this time, NTMS finds that they are the best choices based on our study (See Chapter 5). In fact, HDF/HDF-EOS is the most commonly used data packaging standard in the heritage missions. ESML (Earth Science Markup Language) has the potential to provide a level of data description to enable interchange packaging; however, NTMS finds this technology is not yet sufficiently mature to recommend. There are certainly other choices for standard interchange packaging as well.

NTMS finds that it is unlikely that certain data format standards will fill the needs of a center-to-center data interchange standard. For example, GeoTIFF is a good distribution standard for georeferenced imagery to end-users, but it is incapable of conveying the full metadata needed for the near-term missions or handling non-image data such as atmospheric profiles. BUFR and GRIB are WMO standards designed for dissemination of weather station data and for the output of numerical weather prediction models. BUFR and GRIB may be used as distribution formats for ocean or atmosphere data products used by weather prediction modelers, but they are not suitable as interchange standards as they lack self-describing power, tool support, and other criteria (see Chapter 5).

Finally, some near-term missions have heritage in, or are considering the use of, custom (a.k.a. binary) data formats. We recommend that while it may be appropriate to use custom formats for internal mission science work, and certain communities may find them appropriate for distribution packaging, such formats are unlikely to be acceptable as an interchange standard.

7.3 Metadata Standards Recommendations

1. Metadata standards are converging on the ISO 19115 standard, as FGDC will adopt the ISO 19115 after it becomes final in 2002. We recommend ISO19115 as the metadata standard for the near-term missions. However, since the ISO19115 is not finalized yet and the tool support for it has yet to be developed, we recommend the following implementation strategy.

2. Data systems for near-term missions *must* be compliant with the EOSDIS Clearing HOuse (ECHO) metadata model, which will be used for the advertising and distribution of data from "Backbone" Processing Centers, Science Data Centers, Multi-mission Data Centers, and PI-managed Missions Data Centers. The ECHO metadata model is an XML implementation and a superset of the ECS data model. It provides a capability to map metadata into various content-equivalent representations. With the single investment of ECHO interoperability, near-term missions will be insulated from most metadata standards changes and will benefit from new ECHO interfaces as they become available.

3. The ECHO implementation organization *should* be tasked with monitoring, developing, and maintaining metadata mapping capability between ECHO holdings and emerging FGDC remote sensing profiles of the ISO 19115 standard.

4. We recommend continuing the use of the GCMD as a catalog (or collection) metadata standard for the near-term missions. We further recommend that the GCMD *should* coordinate with ECHO to implement seamless, automated interoperability of data products and services so that data centers do not need to prepare collection (or dataset) level metadata separately from inventory-level metadata.

7.4 Documentation Standards Recommendations

1. The Guide standard *should* be maintained, and a community-based process for incorporating the Guide into the ECHO data model *should* be developed. The EDG Guide data set documentation standard is successful and generally adequate for minimal description of standard data products. However, the division of metadata between the EOSDIS Earth Science Data Model and the EOSDIS Guide Document appears arbitrary and is, we believe, a hindrance to the effective, efficient, and accurate discovery and use of EOS data.

2. Algorithm Theoretical Basis Documents (ATBDs) are written by EOS scientists for every EOS instrument product. There are no detailed specifications for ATBDs, only a suggested outline that includes theoretical background, algorithm description, and validation plans. The ATBD is typically referenced in the Guide document. The ATBDs *should* be permanently accessible with a stable web address so that links to these documents in the Guide documents will remain valid

3. NASA *should* perform a detailed analysis of the emerging XML-based documentation standards in the social and library/archive sciences. The purpose of performing a detailed analysis of the emerging XML-based documentation standards in the social and library/archive sciences is to: 1) borrow from their XML-based data models which combine free-text descriptions and constrained element lists in a hierarchical, cross-referenced fashion; 2) maximize interoperability with information systems outside of Earth Sciences; and 3) benefit from the extensive work that has been done defining the documentation required for long-term preservation of knowledge about digital objects. We are specifically referring to the Data Documentation Initiative (DDI) in the Social Sciences, the Metadata Encoding and Transmission Standard (METS) maintained by the Library of Congress, and the Reference Model for an Open Archival Information System (OAIS) developed by the Council of the Consultative Committee for Space Data Systems.

7.5 Standard Evolution Process and Other Activities Recommendations

1. For Earth systems science systematic measurements including EOS data, weather data, atmospheric and oceanographic modeling data, and land use/land processes data, there are several data packaging or API standards that have similar purposes. ESE *should* invest resources in guiding the evolution of these data formats through their respective governing processes with the goal of harmonizing them toward seamless interoperability. We recommend that particular attention be focused on guiding the evolution of netCDF, HDF/HDF-EOS, geoTIFF and the WMO BUFR and GRIB formats

2. SEEDS *should* empanel a Standards and Interfaces Evolution Process Working Group for developing and executing a plan for evolution of interchange packaging standards over the life of data sets. The benefits of adopting additional interchange packaging standards need to be weighed against the increase in cost and complexity that will occur with the addition of each new packaging standard.

3. Near-term missions *must* plan for evolution of end user requirements for packaging of mission science data (including data distribution packaging formats, data distribution system interface, and metadata) over the lifetime of the missions.

4. The evolution of the interchange packaging standards *must* keep pace with technology and market trends and emerging expectations (see lessons learned from SDTS in Chapter 3). SEEDS *should* adopt new technology as it develops. The number of interchange packaging standards *should* be limited and as closely related to each other as is practical.

5. SEEDS *should* coordinate respective activities to support the near-term missions such as interchange packaging standards maintenance and translation tools development/maintenance. This group *should* also provide interchange packaging standards user training and help desk support to educate producers/consumers/tool vendors.

6. The development of conversion software for data distribution formats *should* be a separately funded task and the responsibility for this development *should not* necessarily fall upon the mission science teams.

Acronym List

ACRIM	Active Cavity Radiometer Irradiance Monitor
ACRIMSAT	Active Cavity Radiometer Irradiance Monitor SATellite
ADEOS	Advanced Earth Observing Satellite
ADS	Archive and Distribution Segment
AGS	Alaska Ground Station
AIRS	Atmospheric Infrared Sounder
AIX	IBM's UNIX Operating System
ALI	Advanced Land Imager
ALT	Dual-Frequency Radar Altimeter
AMI	Active Microwave Instrument
AMSR	Advanced Microwave Scanning Radiometer
AMSU	Advanced Microwave Sounding Unit
ANZLIC	Australia New Zealand Land Information Council
APAS	Astrophysical, Planetary, and Atmospheric Sciences Department
API	Application Platform Interface
APID	Applications Package Identification
ASCI	Accelerated Strategic Computing Initiative
ASPS	AIRS Science Processing System
ASTER	Advanced Spaceborne Thermal Emission And Reflection Radiometer
ATMOS	Atmospheric Observations Satellite
ATMS	Advanced Technology Microwave Sounder
AVHRR	Advanced Very High Resolution Radiometer
AVISO	Validation and Interpretation of Satellites Oceanographic data
AVS	Advanced Visual Systems
BOREAS	Boreal Ecosystem-Atmosphere Study
BSQ	Band Sequential
BUFR	Binary Universal Format For Representation [Of Data]
C3S	Command Control & Communication Segment
CAP	Cooperative Agreements Program
CARS	Climate Analysis and Research
CASE	Computer Aided Software Engineering
CBS	Commission for Basic Systems
CCI	Carbon Cycle Initiative
CCIWG	carbon cycle interagency working group
CCS	Climate Calibration Segment
CDF	Common Data Format
CDHF	Central Data Handling Facility
CDL	Common Data Form Language (used by netCDF

CDMS	Climate Data Management Segment
CDR	Climate Data Record
CEOS	Committee on Earth Observation Satellites
CERES	Clouds and the Earth's Radiant Energy System
CI	Catalog Interoperability
CIP	Catalog Interoperability Protocol
CLAES	Cryogenic Limb Array Etalon Spectrometry
CMIS	Conical Microwave Imager/Sounder
CMS	Climate Mission Storage System
CNES	Centre National D'etudes Spatiales (France)
CNIDR	Clearinghouse for Networked Information Discovery and Retrieval
COTS	Commercial Off-The-Shelf
CPF	Calibration Parameter File
CPOZ	Compressed Ozone
CrIS	Cross-Track Infrared Sounder
CRTT	Calibrated Radiance and Temperature Tape
C-SAFS	Central Standard Autonomous File System
CSDGM	Content Standard for Digital Geospatial Metadata
CZCS	Coastal Zone Color Scanner
DAAC	Distributed Active Archive Center
DAO	Data Assimilation Office
DC	Dublin Core
DEM	Digital Elevation Model
DFD	Deutsches Fernerkundungsdatenzentrum (German Remote Sensing Data Center)
DFR	Dual Frequency Radar
DIAL	Data and Information Access Link
DIF	Directory Interchange Format
DLL	Dynamic Link Library
DLT	Digital Linear Tape
DMF	Data Models and Formats
DMSP	Defense Meteorological Satellite Program
DoD	Department of Defense
DODS	Distributed Oceanographic Data System
DOE	Department of Energy
DOMSAT	Domestic Satellite
DOQQ	Digital Orthophoto Quarter Quads
DORIS	Doppler Orbitography And Radiopositioning Integrated By Satellite
DOS	Disk Operating System
D-PAF	German Processing and Archiving Facility
DPR	Dual-frequency Precipitation Radar
DPS	Data Processing System

DRFP	Draft Request for Proposal
DRG	Digital Raster Graphics
DSP	Directory Service Protocol
DSWG	Data System Working Group
DTD	Document Type Definition
EA	Engineering Analysis
ECHO	EOS Clearing House
ECMWF	European Center for Mid Range-Weather Forecasting
ECS	EOSDIS Core System
EDC	EROS Data Center
EDG	Earth Observing Systems (EOS) Data Gateway
EDOS	EOS Data and Operations System
EDR	Environmental Data Record
EMOS	ECS Mission Operations System
ENSO	El Niño/Southern Oscillation
ENVISAT	Environmental Satellite
EO-1	Earth Orbiting Satellite #1
EOC	EOS Operations Center
EOS	Earth Observing System
EOSAT	Earth Observation Satellite
EOSDIS	Earth Observing System Data and Information System
EP	Earth Probe
EPA	Environmental Protection Agency
EPSG	European Petroleum Survey Group
EROS	Earth Resources Observation System
ERS	Earth Resources Satellite
ERSDAC	Earth Remote Sensing Data Analysis Center (Japan)
ERTS	Earth Resource Technology Satellite (later renamed Landsat 1 (Land Saltellite?)
ESA	European Space Agency
ESC	Engineering Support Center
ESCAT	ESA scatterometer
ESDIS	Earth Science Data and Information System
ESDT	Earth Sciences Data Type
ESE	Earth Science Enterprise
ESIP	Earth Science Information Partner
ESIPS	EOSDIS Science Investigator-Led Processing System
ESRI	Environmental Systems Research Institute (GIS software company)
ETM	Enhanced Thematic Mapper
EUV	Extreme Ultraviolet
FAST-L7A	FAST-Landsat 7 Format
FDF	Flight Dynamics Facility

FGDC	Federal Geographic Data Committee
FIFE	First ISLSCP Field Experiment
FIPS	Federal Information Processing Standard
FMI	Finnish Meteorological Institute
FTP	File Transfer Protocol
FX	File Transfer Subsystem
GAC	Global-Area Coverage
GCCP	Global Carbon Cycle Program
GCMD	Global Change Master Directory
GCTE	Global Change and Terrestrial Ecosystems
GDAAC	Goddard Distributed Active Archive Center
GDAL	Geospatial Data Abstraction Library
GDR	Geophysical Data Record
GEO profile	Geospatial Metadata Application
GeoTIFF	Georeferenced Tagged Image File Format
GES	Goddard's Earth Sciences
GHRC	Global Hydrology Resource Center
GILS	Global Information Locator Service
GIS	Geographic Information System
GIVIT	Granule Insert Validation and Inspection Tool
GLAS	Geoscience Laser Altimeter System
GLOBEC	Global Ocean Ecosystem Dynamics
GMT	Greenwich Mean Time
GOES	Geostationary Operational Environmental Satellites
GOFC	Global Observation of Forest Cover
GOME	Global Ozone Monitoring Experiment
GPM	Global Precipitation Measurement
GRACE	Gravity Recovery And Climate Experiment
GRASS	Geographic Resources Analysis Support System (public domain software)
GRiB	GRidded Binary
GRS-1	Generic Record Syntax
GSFC	Goddard Space Flight Center
GUI	Graphical User Interface
HAO	High Altitude Observatory
HDF	Hierarchical Data Format
HDF-EOS	HDF Earth Observing System (EOS) format
HDSBUV	High-Density Solar Backscatter Ultraviolet Instrument (SBUV)
HE4	HDF-EOS based on HDF version 4
HE5	HDF-EOS based on HDF version 5
HIRDLS	High-Resolution Dynamics Limb Sounder
HRDLS	High-Resolution Dynamics Limb Sounder

HRPT	High Resolution Picture Transmission
HSB	Humidity Sensor of Brazil
HTTPD	Hyper Text Transfer Protocol Daemon
IAS	Image Assessment System
ICD	Interface Control Document
ICESat	Ice, Cloud, And Land Elevation Satellite
IDL	Interactive Display Language
IDN	International Directory Network
IDPS	Interface Data Processing System
IETF	Internet Engineering Task Force
IFOV	instantaneous field of view
IGBP	International Geosphere And Biosphere Research Program
IGDR	Interim Geophysical Data Record
IGS	International Ground Stations
IMS	Information Management System
IPD	Information Processing Division
IPO	Integrated Program Office
ISCCP	International Satellite Cloud Climatology Project
ISLSCP	International Satellite Land Surface Climatology Project
ISO	Greek prefix "iso" as used by the International Organization for Standardization
ITSS	Information Technology and Scientific Services
IWGDMGC	Interagency Working Group on Data Management for Global Change
JAI	Java Advanced Imaging
JEB	Java EOS Browser
JERS	Japanese Earth Resources Satellite
JMR	Jason Microwave Radiometer
JPL	Jet Propulsion Laboratory
KLM	NOAA K-, L-, M- system
KNMI	Koninklijk Nederlands Meteorologisch Instituut (Netherlands)
LAC	local-area coverage
LaRC	Langley Research Center
LASP	Laboratory for Atmospheric and Space Physics at the University of Colorado
LBA	Large-Scale Biosphere-Atmosphere Experiment
LDCM	Landsat Data Continuity Mission
LGN	Landsat Ground Network
LGS	Landsat Ground Station
LIS	Lightning Imaging Sensor
LP	Level Processor
LPDS	Level 1 Product Distribution System
LPGS	Level 1 Product Generation System
LPS	Landsat Processing System

LTER	Long-Term Ecological Research
MBLA	Multi-Beam Laser Altimeter
MCF	Metadata Configuration File
METI	Ministry Of Economy Trade And Industry (Japan)
MFLOPS	Millions of Floating Point Operations per Second
MINC	Medical Image netCDF
MISR	Multi-Angle Imaging Spectro-Radiometer
MIT	Massachusetts Institute of Technology
MLE	Maximum Likelihood Estimator
MLS	Microwave Limb Sounder
MOBY	Marine Optical Buoy data
MOC	Missions Operations Center
MODAPS	MODIS Adaptive Processing System
MODIS	Moderate-Resolution Imaging Spectroradiometer
MOPITT	Measurements Of Pollution In The Troposphere
MSCD	Mirror Scan Correction Data
MSFC	Marhsall Space Flight Center
MSS	Multispectral Scanners
MTMGW	Machine-To-Machine Search and Order Gateway
MUSE	Multi-User Science Environment
NASA	National Aeronautics and Space Administration
NASDA	National Space Development Agency (Japan)
NCAR	National Center for Atmospheric Research
NCDC	National Climatic Data Center
NCEP	National Centers for Environmental Prediction
NCL	NCAR Command Language
NCSA	National Center for Supercomputing Applications
NDVI	Normalized Differential Vegetation Index
NESDIS	NOAA/National Environmental Satellite, Data, and Information Service
netCDF	Network Common Data Format
NIVR	Netherlands's Agency for Aerospace Programs
NMC	National Meteorological Center
NOAA	National Oceanic and Atmospheric Administration
NODC	NOAA National Oceanographic Data Center
NPOESS	National Polar Orbiting Environmental Satellite System
NPP	NPOESS Preparatory Project
NRL	Naval Research Laboratory
NRT	Near Real-Time
NSCAT	NASA scatterometer
NSF	National Science Foundation
NSIDC	National Snow and Ice Data Center

NSSDC	National Space Science Data Center
NTMS	Near-Term Missions Standards
NTMS	Near-Term Missions Study
NWP	Numerical Weather Product
NWS	National Weather Service
ODL	Object Description Language
OFL	Off-Line
OGC	OpenGIS Consortium
OMB	Office of Management and Budget
OMI	Ozone Mapping Instrument
OMPS	Ozone Mapping and Profiler Suite
ORNL	Oak Ridge National Labs
OSDPD	Office of Satellite Data Processing and Distribution
OSDR	Operational Sensor Date Record
OSF	Observation Schedule File
OSSIM	Open Source Software Image Map
OSTM	Ocean Surface Topography Measurement
OTTER	Oregon Transect Ecosystem Research
PCD	Payload Correction Data
PCF	Process Control File
PCI	GIS software by PCI Geomatics
PDPS	Planning and Data Processing System
PDR	Product Delivery Record
PDS	Precipitation Data System
PGE	Product Generation Executive
PI	Principle Investigator
PMEL	Pacific Marine Environmental Laboratory
PNG	Portable Network Graphics
PO.DAAC	Physical Oceanography DAAC
POES	Polar-Orbiting Environmental Satellite
POSC	Petrotechnical Open Software Corporation
PP	Pre-Processors
PPLUS	graphics package by Plot Plus Graphics
PR	Precipitation Radar
PSA	Product Specific Attribute
PV-wave	Visualization package by Visual Numerics, Inc.
QAMUT	Quality Assurance Metadata Update Tool
QuickScat	Quick Scatterometer
QuikTOMS	Quick TOMS
RBV	Return-Beam Vidicons
RDBMS	relational database management system

RDF	Resource Description Framework
RDR	Raw Data Record
RFP	Request for Proposal
RGB	Red, Green, Blue
RSI	Research Systems, Inc.
RSS	Remote Sensing Systems
S4P	Scalable Script-Based Science Processor
SA	Science Analysis
SAA	Satellite Active Archive
SAGE	Stratospheric Aerosols and Gas Experiment
SAIC	Science Applications International Corporation
SAN	Storage Area Network
SAR	Synthetic Aperture Radar
SASS	Seasat-A scatterometer system
SBUV	Solar Backscatter Ultraviolet Instrument
SCF	Science Computing Facility
SCIAMACHY	SCanning Imaging Absorption SpectroMeter for Atmospheric ChartographY
SCLI	Science data server Command-Line Interface
SDP	Science Data Processing
SDPS	SeaWifs Data Processing System
SDS	Science Data Segment
SDSRV	Science Data Server
SDTS	Spatial Data Transfer Standard
SeaPAC	SeaWinds Processing and Analysis Center
SeaWiFs	Sea-viewing Wide Field-of-view Sensor
SEE	Solar EUV Experiment
SEEDS	Strategic Evolution of ESE science Data and information Systems
SERF	Service Entry Resource Format
SFDU	Standard Formatted Data Unit
SGDR	Sensor Geophysical Data Record
SGI	Silicon Graphics, Inc.
SGML	Standard Generalized Markup Language
SGS	Svalbard Norway Ground Station
SIGF	Solar Irradiance Gap Filler
SIM	Spectral Irradiance Monitor
SIPS	Science Investigator Processing System
SMI	Standard Mapped Image
SMMR	Scanning Multichannel Microwave Radiometer
SMMS	Spatial Metadata Management System
SNOE	Student Nitrous Oxide Experiment
SOLSTICE	Solar Stellar Irradiance Comparison Experiment

SORCE	SOlar Radiation and Climate Experiment
SPOT	Systeme Pour l'Observation De La Terre (France)
SPS	Science Processing System
SQL	Structured Query Language
SRM	Subscription Request Manager
SRTM	Shuttle Radar Topography Mission
SSALT	Single-Frequency Solid-State Radar Altimeter
SSM/I	Special Sensor Microwave/Imager
SUTRS	Simple Unstructured Text Record Syntax
SWIR	Short Wave Infrared
TBD	To Be Determined
TDI	Transport Data Interface
TES	Tropospheric Emission Spectrometer
TIFF	Tagged Image File Format
TIM	Total Irradiance Monitor
TIMED	Thermosphere Ionosphere Mesosphere Energetics and Dynamics
TIROS	Television Infrared Observation Satellite
TM	Thematic Mapper
TMI	TRMM Microwave Imager
TMR	TOPEX Microwave Radiometer
TOMS	Total Ozone Mapping Spectrometer
TOPEX	Topography (Ocean) Experiment
TOVS	Tiros Operational Vertical Sounder
TRMM	Tropical Rainfall Measurement Mission
TSDIS	TRMM Science Data and Information System
TSIS	Total Solar Irradiance Sensor
TSU	TSDIS Science User
UARS	Upper Atmospheric Research Satellite
UCSS	UARS CDHF Software System
UDDI	Universal Description, Discovery, and Integration
UFM	User Friendly Metadata
UML	Universal Modeling Language
USDA	United States Department of Agriculture
USFS	USDA Forest Service
USGCRP	US Global Change Research Program
USGS	United States Geological Survey
USMARC	U.S. Machine Readable Cataloging
UTM	Universal Transverse Mercator
VCL	Vegetation Canopy Lidar
VDC	Visual Database Cookbook
VIIRS	Visible/Infrared Imager/Radiometer Suite

VIRS	Visible and Infrared Scanner
VNIR	Visible And Near Infrared
W3C	World Wide Web Consortium
WAIS	Wide Area Information System
WGISS	Working Group on Information Systems and Services
WHO	World Health Organization
WIM	Windows Image Manager
WMO	World Meteorological Organization
WNS	Wind Scatterometer
WOCE	World Ocean Circulation Experiment
WRS	Worldwide Reference System
WXP	Weather Processor
XDR	eXternal Data Representation
XML	eXtensible Markup Language
XPS	XUV Photometer System
XSL	eXtensible Stylesheet Language
XSLT	XSL Transformations
ZMT	Zonal Mean Tape

Appendix C – Metrics Planning and Reporting

Section 6 of "SEEDS Metrics Planning and Reporting: Accountability Survey Report."

6 Summary of Results and Conclusions

This Section summarizes preliminary results, draws initial conclusions, etc., from the survey responses received so far, addressing organizational relationships and funding mechanisms, metrics collection and reporting, accountability, general issues, and overall results and conclusions. This report is based on responses received from eighteen activities of the thirty that were importuned (two additional partial responses were received).

Of the eighteen sites that responded, five are DAACs, ten are ESIPs, one is a SIPS, one (DODS) is a distributed data transport system, and one is a NASA space science data center. Of the ten ESIPs, five are science data centers (ESIP type 2), four are applications activities (ESIP type 3), and one is a data center (ESIP type 1). (Technically, the DAACs are also regarded as Type 1 ESIPs.)

The mix of responding activity types is:

> 7 Data Centers (LP DAAC, PO.DAAC, ORNL DAAC, GES DAAC, NSSDC, GHRC, SEDAC)
> 1 Processing Center (AMSR-E SIPS)
> 5 Science Data Centers (GLCF, SIESIP, EOS-WEBSTER, OceanESIP, PM-ESIP)
> 4 Applications Activities (EDDC, TerraSIP, BASIC, TERC)
> 1 Infrastructure Activity (DODS)

6.1 Organizational Relationships and Funding Mechanisms

One objective of the study was to learn what administrative and funding mechanisms are used to fund ESE activities and assess the appropriateness of them given the nature of the activity and the degree of satisfaction with them by both activities and their sponsors, and to see if activities funded by more than one sponsor experienced conflicts as a result. Another objective was to understand the institutional commitment of each activity's host organization to its activities. This preliminary report addresses only the activity point of view.

All eighteen responding activities are entirely or mostly funded by NASA, though often not by a single NASA sponsor.

Three responding activities are funded on contracts, ten activities are funded on cooperative agreements, two activities are funded by grants, two activities are funded through NASA's internal processes, and one activity is funded through an interagency agreement.

Eleven of the eighteen responding activities report satisfaction with the funding mechanism under which they currently operate. Three activities funded under cooperative agreements reported difficulties: one that the terms of the cooperative agreement conflicted with the NASA contract under which its host institution operates, one with the promptness of NASA's payments, and one with terms of the cooperative agreement that prevented it from partnering with a private company. The remaining four activities, two that operate on cooperative agreements and two that operate on NASA's internal funding processes, did not comment.

As an additional measure of their satisfaction, fourteen of the eighteen responding activities reported that they had the authority they needed to carry out the responsibilities assigned by their sponsors. One activity reported that NASA's failure to make payments promptly undercut the ability of the activity to manage its work effectively. Another activity reported difficulty with its manager being held responsible for metrics embracing projects not under the manager's control. A third activity reported that, as a SIPS, it would like to have the authority to distribute data to some near real-time users but could not. A fourth activity reported difficulty with long lead times required for foreign travel approval and restrictions on equipment purchase authority.

In addition, two DAACs noted as a soon-to-be-relieved exception a lack of authority over the EOSDIS Core System at their sites, and one mentioned difficulty in resolving inconsistencies in interpretation of policies (e.g. user privacy, website content). The exceptions reflect unique circumstances at the activities involved rather than a more fundamental problem traceable to the funding or administrative mechanisms under which they operate.

Activities with a primarily operational function supporting the ESE program (e.g. DAACs) were funded by contract, interagency agreement, or NASA's internal processes, with two exceptions: a SIPS and a data center were funded under a cooperative agreement. Activities with a primarily research (e.g. science data centers) or applications development mission were funded by cooperative agreements or a grant. In all but two cases the administrative and funding mechanisms were appropriate given the missions of the activities.

The two cases in question involved operational activities performed under a cooperative agreement. NASA procurement guidance (the "principal purpose of the relationship" clause found in NPG 5800.1, Part 1260.12, Choice of award instrument, Grant and Cooperative Agreement Handbook) states that the choice of award instrument should follow a principal purpose test, and that "if the principal purpose of a transaction is to accomplish a NASA requirement, i.e. to produce something for NASA's own use, a contract is the instrument to be used. Operational SIPS and data center functions seem to meet the criterion for a contract.

Thirteen of the eighteen responding activities report receiving funding from more than one source, with twelve having no difficulty with conflicts between funding sources and one not reporting. One activity notes the possibility of future conflicts when ESE data management is funded through flight projects with no responsibility for data stewardship beyond the duration of the flight mission. Several noted that potential conflicts have been successfully avoided by constructive discussions with sponsors. Five activities report having a single funding source, hence no possibility of conflict.

Nine activities report some degree of satisfaction with contracts as a mechanism for funding operating elements (some internal, such as an on-site support contractor, some distributed). One reports a difficulty with university subcontracts, because the university expects to do research, and the requirement was for delivery of products. One reported that problems with prompt payment by NASA created problems with payments to subcontractors. Two reported difficulty with a cumbersome procurement process of their host institution. The other nine activities did not report funding internal elements

One activity relies on a DAAC for part of its support, and notes that it does not fund the DAAC for the work the DAAC does, and that its arrangement with the DAAC "requires a higher degree of management skills to maintain program effectiveness and a continuing effort to ensure that they DAAC receives its approved support".

6.2 Metrics Collection and Reporting

An objective of the study is to obtain an understanding of the usefulness of the metrics that are currently being used, any problems that activities are having with them, and any recommendations activities have for improving them in the future.

General Response on Metrics:

The DAACs and ESIPs that responded to the survey provide metrics required by NASA, including overall GPRA metrics defined by NASA Headquarters and more detailed metrics required of DAACs by the ESDIS Project and of ESIPs by their NASA Headquarters sponsor.

Most (twelve, with one not reporting, and two that were not ESIPs when ESIP metrics were defined) of the responding activities thought they had had sufficient input into definition of the metrics that they were required by their sponsor to collect and report, though three DAAC responses were qualified.

There was a significant difference in the times that the data centers reported they spend on metrics (from 0.25 FTE to between 1 and 2 FTE per year) reporting weekly and monthly, and the science data centers (from 4 days per year to 24 days per year or 0.02 to 0.1 FTE per year) on metrics reporting quarterly. When the science data centers begin reporting on a monthly basis, the effort level will rise.

All of the responding activities agreed that the metrics they were required to report were generally useful, but there was agreement that the metrics in themselves don't provide a complete picture, e.g. they don't measure that value of an activity's data and services to its user community or sponsor, and they don't measure effort in research work, training, outreach, etc. The ESIPs that responded highlighted the importance of the 'nuggets' that provide information on user satisfaction that is not captured by quantitative metrics.

Metrics Most Useful for Internal Management:

Activities pointed to quantitative measures showing growth (in user community, products delivered, etc.), measures that provided guidance for performance tuning, that measured the effectiveness of specific operations, system enhancements, utilization of data sets, utilization of system resources in key areas (e.g. data production and distribution), system availability, production actual vs plan, peak demand on computers/networks including concurrent users. Several activities report collection of user feedback (e.g. user email, user logs, user comments) for identification, analysis and resolution of performance or service problems. One activity reports the analysis of orders cancelled or delayed to resolve the problems involved.

Metrics and How an Activity's Success is Judged by Sponsor and User Community:

All responding activities reported that users are primarily concerned with getting good service (e.g. easy access to readily usable data, tools, effective user support), and they judge activities on that basis. Users are seen as indifferent to aggregate metrics - being understandably concerned with the service they personally receive.

Sponsors are seen as having to some degree different concerns from users; while users want good service sponsors want to see that the activity is important to ESE, that users seek its data and services. Sponsors are seen as judging activities by a combination of feedback from users and metrics. If individual users feel well served (or not), feedback will informally percolate up to the sponsors, in the case of the DAACs often via User Working Groups whose meetings are attended by ESDIS and NASA Headquarters. Other avenues exist for feedback from flight projects or instrument teams. This feedback (which was noted sometimes gives an activity credit or blame in an area for which it is not responsible) is seen as carrying great weight with sponsors, but is not entirely reflected in the quantitative metrics. ESIPs note that the 'nuggets' that they provide are the best measures of user satisfaction.

One activity stressed the need for the sponsor to request independent feedback from the users of sponsored activities in order to make their own assessment of the performance of an activity without depending solely on information provided by the activity. Another did not favor sponsors going directly to users asking about their satisfaction, suggesting that the activity will get the most complete feedback and can relay that feedback back to the sponsor.

DAACs and ESIPs generally pointed to a need to develop metrics that measure the value of data to scientists and their use of data, mentioning citations in peer reviewed literature as a possibility. One DAAC (SEDAC) reports that collection of citations of its products and services in scientific and technical literature, published documents and reports, and on-line material is more helpful than other metrics in determining the overall utility of its data and products, though noting that there is often a substantial delay before the citations appear. NSSDC also reports that citations are a key measure of success for its work, more than reported statistics.

Problems Reported with Metrics:

There was agreement that the sponsor-required DAAC or ESIP quantitative metrics do not capture user satisfaction, and do not measure the value of an activity's data and services to the science or applications community, or the actual utilization of data by the communities. One activity emphasized that the metrics often only reflect the user's experience with locating and obtaining data, and not their experience in understanding the data and in applying it to meet their needs. Another activity noted that simple metrics such as volume distributed will not reflect the users' satisfaction when services such as subsetting are more extensively used, which could result in decreased volume delivered (at greater effort by that activity) but greater user satisfaction.

DAACs reported problems with inconsistencies between their own metrics and what they see reported from the ESDIS Project's metrics system that draws on inputs from the DAACs. DAACs also complained that metrics are misinterpreted - and that they are never consulted about how the metrics ought to be interpreted.

One ESIP noted that ESIP Federation metrics are often incomplete and inaccurate, but other ESIPs did not make the same comment. One ESIP noted that the defined ESIP metrics did not treat small activities fairly in comparison with larger activities. Another ESIP pointed out the difficulty in applying the ESIP metrics to the multiple projects comprising its activity.

There was a general subtext that insufficient resources were provided by sponsors to meet increasing requirements for metrics. The sponsors added new metrics beyond the set that activities had helped to define, without adding resource support.

Recommended Changes to Metrics:

There is a consensus that measures of the utilization of activities' data and service, and the value of those data and services, to the science or applications community are needed. Citations in peer reviewed literature were suggested as one such measure, and indeed SEDAC and NSSDC report that they rely on citations as a key measure. One ESIP noted that the current metrics are 'data centric', and suggested that what are needed are measures of the promotion of the conduct of science, such as the number of publications about data activities in peer reviewed journals, graduate students attracted to programs promoted by an activity, students graduated from Earths science programs, etc., in addition to citations. Another suggestion was to measure value by broadening of the user base for an activity's data and services, including tracking outreach activities and the attendance at them. (It was noted in the course of a Workshop discussion that publications tend to lag behind data by a year or two; a comment centered around getting outcome metrics rather than output metrics. In this light, publications can be used as a metric for data generated 2 years ago. In other words, it's now time to start looking for publications related to Terra data.)

One suggestion was that a portion (e.g. 25%) of the metrics reported to its sponsor by an activity be selected by the activity to be part of the basis for the sponsor's evaluation of the activity. This would allow an activity to include metrics appropriate to its particular nature or circumstances in addition to metrics applied to all activities. Another activity noted that the standard set of ESIP metrics did not reflect its efforts in starting and completing a number of small projects with the education community. If this idea were implemented, that activity could add specific metrics reflecting that work (e.g. number of projects, number of school districts and teachers involved).

Other suggestions included metrics that would reflect the number and type of special processing and services provided (e.g. subsetting, remapping, reformatting), the number of users requesting such services, effect of special services on data volume, etc., and effort required.

Other suggestions for improved metrics were measures of how effective an activity is in meeting ESE data management needs, and measures of integration/development support provided by an activity.

Recommendations on 'What could a SEEDS Office Do' regarding Metrics or for New Means of Publicizing Accomplishments:

One activity recommended: "Development of a systematic, cross-DAAC search for citations and data usage in the scientific, policy, and popular literature and in online information resources. Such an effort would be more cost effective and less subject to bias if conducted for all DAACs by a third party such as a SEEDS office. The "hits" from such a search could be tabulated quantitatively and be used as the basis for documenting significant uses of data, e.g., in an important scientific publication or significant policy decision. Such materials could then be used by the NASA Earth Observatory, the DAAC Alliance Yearbook, and other outreach efforts."

Another activity noted that a 'SEEDS Office' could require ESE activities to identify papers that highlight or use their products and collect them periodically into special volumes. A 'SEEDS Office' could publish an annual report that includes a brief summary of the work of each ESE activity, plus the first page of key papers published that were based on the activity's data.

One activity noted that metrics have been usually defined afterwards rather than before. As a result, it has been difficult to go back and regenerate new metrics. A 'SEEDS Office' through this and other surveys could anticipate the metrics desired and/or required by policy makers, HQ management, and lead center technical management.

ESIPS suggested that a 'SEEDS Office' could help the Federation advertise on NASA websites, could coordinate / facilitate outreach activities by activities, e.g. attendance at conferences or workshops by appropriate activities. A 'SEEDS Office' could sponsor a special journal issue or articles in journals about ESE activities, it could organize a conference focusing on research contributions made by activities, it could organize workshops highlighting data tools and products developed by activities. A 'SEEDS Office' could facilitate partnering with other NASA programs such as the Space Grant Education program and other educational activities. A 'SEEDS Office' could proactively link NRA's, AO's, etc., with ESIP activities, emphasizing the data and services provided by or available from the ESIPs.

6.3 Accountability

An objective of the study is to understand the requirements levied on activities for accountability in selected areas, and ideas activities have on improving the processes by which such requirements are enforced, keeping in mind that the requirements should be commensurate with an activity's functions.

Accountability Requirements for IT Security:

There was a marked difference between the data centers (DAACs and the other Type 1 ESIP) with NSSDC not commenting) and the science data centers and applications activities (ESIPs). All of the DAACs described compliance with NASA sponsor and Federal IT security standards and practices, with one reporting the added burden of reporting to both its host institution and ESDIS, driven by the same NASA requirement. Although ESIP host institutions have their own IT security practices, and the ESIP activities themselves instituted practices they regarded as sensible, they reported no sponsor-driven requirements for IT security or related reporting.

Other Accountability Requirements (e.g. user privacy, web-site accessibility):

The DAACs are required to follow NASA sponsor and government wide mandates on user privacy and website accessibility, with reporting on compliance mainly ad hoc as requested by ESDIS, or through regular means (e.g. weekly telecons, weekly and monthly reports).

ESIPs variously report a requirement to register products with GCMD, one follows DAAC user privacy guidelines, one guarantees confidentiality of user information, one notes that user privacy is protected but not reported on, others did not have requirements but instituted practices they deemed sensible. In some cases an ESIP's host institution imposed requirements.

Accountability for Data Stewardship:

None of the responding activities reported any accountability requirements or reporting for data stewardship, beyond the routine system metrics they provide. The DAACs feel accountable for data stewardship, to the user community and NASA, and one noted accountability to its User Working Group where stewardship issues are discussed and to which the DAAC responds. One ESIP reported that it takes measures to protect data that proves to be of continuing value to the science community, two others that they are required to pass its data to a DAAC if it goes out of operation.

Recommended Improvements to Accountability Mechanisms:

One activity recommended a common set of requirements that is better defined by the sponsor. Another recommended that better tools be developed for coordination of reporting between distributed data providers - especially by SIPS and science data centers. A third activity, noting the many and diverse uncoordinated requirements stemming from sponsors' interest, federal laws, and other sources, suggested improved coordination and streamlining of reporting requirements to reduce the administrative burden they pose.

6.4 Conclusions

The following are conclusions reached from the sample of eighteen responding activities. Although the sample of eighteen responding sites is not large, it represents a fair cross section of the population of data activities, including data centers (e.g. DAACs and NSSDC), an operational science processing center (a SIPS), science data centers (ESIP-2s), applications activities (ESIP-3s), and one infrastructure activity (DODS). These results do not include any input from the sponsor side of the sponsor-activity relationship.

1. The current use of administrative and funding mechanisms are mostly appropriate and successful.

The activities operate under a variety of administrative and funding mechanisms (contract, cooperative agreement, grant, internal NASA process, inter-agency agreement). In most cases (sixteen of eighteen) the funding mechanism was appropriate to the mission / role of the activity. There were two cases of an operational activity funded under a cooperative agreement that seemed inconsistent with NASA guidance. In a majority of cases (eleven of eighteen, with four not commenting) the activity was satisfied with the mechanism. Three activities reported difficulties with the mechanism (see above). In most cases (fourteen of eighteen) activities felt that they had the authority they needed, while four activities described problems that they felt compromised their authority.

There was also no difficulty reported by any activity in handling conflicts arising from multiple sponsors.

2. Sponsor Required Metrics are Useful, but Miss User Satisfaction and Value to Users.

Seventeen of the eighteen responding activities are ESE data activities (five DAACs, another Type 1 ESIP, a SIPS, and ten ESIPs) who provide metrics required by the ESDIS Project and NASA Headquarters. Both groups felt that while statistics (e.g. production, distribution, archive growth, etc.) provided some useful information to sponsors, they did not include measures of factors by which users judged activity success - providing easy access to readily usable, well-supported data, products, and services. They provided no measure of the value of the activities' data and services to users. The one partial exception noted by the ESIPs was the 'nuggets', user anecdotes, that convey feedback from particular users. Some approaches to measures of utilization and value were suggested, such as tracking citations in peer-reviewed literature, which is currently seen as a key measure by one DAAC and the one non-ESE activity that responded, NSSDC.

3. Future Role of a 'SEEDS Office' Could Include Improve the Measure of User Satisfaction

An ESE level 'SEEDS Office' could develop a systematic, cross-DAAC, and cross-ESIP as appropriate, search for citations and data usage in the scientific, policy, and popular literature. As one activity pointed out, "such an effort would be more cost effective and less subject to bias if conducted for all DAACs by a third party such as a SEEDS office. The results of such a search could be used as the basis for documenting significant uses of data, e.g., in an important scientific publication or significant policy decision, advancing the ESE science and applications program. A 'SEEDS Office,' could also require and fund ESE activities to assemble periodic special collections of papers based on their data and products.

4. The Question of Accountability Needs Study and Policy Review

The responses of the activities to the accountability questions revealed a wide disparity between accountability requirements and reporting between the data centers and the other activities. While some difference might be expected given the operational nature of the data centers, the groups seem to operate at two extremes, strict requirements and reporting by the data centers on one hand, and virtually nothing on the other, with the performance of ESIPs in IT security, user privacy, etc., being a matter of host institution practices and their own judgment. Looking ahead to the SEEDS area, a review of accountability policies seems in order.

5. Accountability for Data Stewardship is a Special Case Needing Study

The responses of the activities to the data stewardship question suggest the need for a review of what the role of a ‘SEEDS Office’ or other ESE program office should be in accepting responsibility for data stewardship across ESE and under what framework of policies or guidelines for practices that responsibility should be delegated out to the ESE data activities. The activities, especially the data centers as perhaps would be expected, are aware of their responsibility for data stewardship, in some cases pointing to their User Working Groups as focal points for their attention. But they do not report any guidelines or requirements from, or reporting back to, their sponsors on data stewardship, and in one case suggested that a trend towards funding data management through flight projects, by their nature temporary, would undercut support for data stewardship. A review of data management planning, data stewardship practices, and metrics that would measure success or detect problems, seems needed.

Section 3 of "SEEDS Metrics Planning and Reporting" draft final report.

3.0 Levels of Accountability

This section defines five attributes that are primary contributors to defining the degree, or level, of accountability for SEEDS data services providers. The attributes are: timeliness, accessibility, dependencies, product quality and data maintenance. Each attribute is associated with a requirement or level of service within a functional area (e.g., ingest, processing, access and distribution) of the Data Service Provider Reference Model as defined in *SEEDS Working Paper 5: Requirements and Levels of Service*. It is not unusual for an attribute to appear in more than one functional area, as does timeliness in the three functional areas mentioned above.

The five attributes are not exhaustive, but represent a core set of requirements that are broad enough to characterize both functionality and accountability of a SEEDS data service provider. The goal is to apply the five attributes to potential SEEDS data service providers and to identify their degree of accountability, or level of accountability, and associate this classification with appropriate funding mechanisms, metrics collection and monitoring mechanisms, and governance. Where appropriate, SEEDS-related language will be recommended for inclusion into various ESE solicitation opportunities.

To differentiate various levels of accountability, three levels are defined for each attribute and are described in the tables below. The three levels are high, medium and low, and are based primarily on the data service provider requirements and levels of service for 14 functional areas described in WP5. WP5 actually presents two views of requirements and levels of service: a technical, detailed listing, and a user-oriented view, i.e., what a user sees in the various levels of service and performance that a data service provider offers. The examples cited below are also from WP5.

For classification purposes, it is recommended that a data service provider be classified at its highest level of accountability for any of the five attributes. For example, if only one of the five attributes is applicable, the SEEDS data service provider will carry a High level of accountability that will affect its funding mechanisms, metrics collection and reporting mechanisms and governance.

Finally, the accountability classification presented here is not intended to be a strict "black or white" or and "easy-to-bin" scheme. The information is intended to be a guideline. Variations will occur and personal judgment will be necessary.

3.1 Timeliness

Timeliness is a critical attribute for a data service provider that appears in ingest, processing, and access and distribution functional areas. An example of this attribute appears in processing as the following level of service: "Operational products shall be generated within 2 days of ingest / availability of required inputs."

TABLE 3.1 Timeliness Accountability Levels

Accountability Level	Timeliness Requirement	Description
High	Time-critical, schedule driven operations	All operations schedule-driven; near-real-time critical time constraints; all events scheduled. On-demand production with time constraints. Impact of an operational problem likely to be severe.
Medium	Non-time-critical, scheduled operations	Operations nominally scheduled; time constraints are not critical; non-real-time events. While impact of a problem can be severe, there is more leeway for resolution before criticality.
Low	Ad hoc, intermittent; schedule not critical	Unscheduled, non-real-time events. Impact of a problem is unlikely to be severe.

3.2 Accessibility

Accessibility appears primarily in search and order, access and distribution, and user support functional areas. Accessibility is a critical attribute that directly supports the availability of ESE data and products. Examples of this attribute are: "Public access to all users for search and order, and access / distribution" and "Help desk staffed five days a week, 8-hours per day." Accessibility can also be seen as a measure of the number of users that would be impacted by a data services provider problem (though some requirements to support a small user group may be very stringent, and problems can provoke intense responses).

TABLE 3.2 Accessibility Accountability Levels

Accountability Level	Accessibility Requirement	Description
High	Search and order, data, products and services' including user support, are public, open to all users	Services must support large, heterogeneous user community (on the order 10,000 - 100,000), high number of interactions. Problems have wide public exposure.
Medium	Search and order, data, products and services, including user support, are available to the science and	Services focused on science and applications users (on the order of 1,000 - 10,000), can assume users have

	applications community	science background. Problems more contained.
Low	Search and order, data, products and services, including user support, are available to a limited team of scientists or applications specialists	Services can be customized to meet needs of small, homogeneous group of users (on the order of 20 - 100). Problems affect only this small group.

3.3 Dependency

Dependency is when a (source) data service provider is required to provide data, products or services to another (recipient) data service provider. The recipient data service provider typically depends on the source's data, products, or services to perform some or all of its functional areas (e.g., processing, distribution, and archive).

TABLE 3.3 Dependency Accountability Levels

Accountability Level	**Dependency Requirement**	**Description**
High	Requires ingest of satellite data streams for product processing; and creates and distributes products required by other data service providers	Ingest of Level 0, or similar satellite data streams; others depend critically on receiving your product(s) in order to perform their functions; performed on an scheduled, operational basis
Medium	Creates and distributes products for use by other data service providers	Others depend on receiving your product in order to perform their functions; could be operational or non-operational
Low	Creates products, but others do not depend on receiving them	Others do not depend on receiving products from you

3.4 Product Quality

Product quality pertains to standard product generation, and is a critical attribute that appears primarily in processing and documentation functional areas. Several examples are: "The data service provider shall accept standard, research product generation software, and/or data integration and data mining software from users" and "Data and product holdings (including multiple versions of products and corresponding documentation as needed) documented to ESE / SEEDS adopted standard for long-term archiving, including details of processing algorithms, processing history, etc."

TABLE 3.4 Product Quality Accountability Levels

Accountability Level	**Product Quality Requirements**	**Description**
High	Products generated with peer-reviewed science algorithms; validated, provisional and beta data production supported; robust documentation, quality parameters flagged	Standard products used by users who require science-quality products in their processing and analyses.
Medium	Variable product quality; quality parameters flagged	Ad-hoc products used primarily by science team
Low	Quality unknown; documentation minimal or doesn't exist	Experimental products, use at own risk

3.5 Data Maintenance

Data maintenance involves long-term archive of data and products. The attribute appears in the archive functional area of and an example is "Archive capacity is cumulative of all data ingested plus all products generated."

TABLE 3.5 Data Maintenance Accountability Levels

Accountability Level	**Data Maintenance Requirements**	**Description**
High	Long-term data stewardship of Level 0 and higher data products received and generated at a DSP	Applicable to long-term data archival facilities where ongoing stewardship is critical to preserving science value of data
Medium	Pre-determined data sets and / or storage capacity limited by a specified threshold	Applicable to local working storage only, data sets may be separately archived or there may be a short-term urgency for stewardship until data sets go to archive.
Low	Temporary or local working storage	Interim data and products; not for archive

Appendix D – Reference Architecture and Reuse

SEEDS Reuse & Reference Architecture Study: Assessment of Approaches and Processes

Contract NAS5-99001
Deliverable 3
File “SEEDS reuse study summary v1.0.doc”
3 July, 2003

Abstract

Meeting the goals of NASA’s Earth Science Enterprise over the next ten years will require new approaches to developing Earth science data systems. How can we deliver new capabilities to meet science and application needs within constrained budgets and schedules? One possible answer is software reuse: by better leveraging existing data system assets, NASA’s ESE should be able to focus its efforts on new capabilities rather than “reinventing the wheel”.

The Strategic Evolution of ESE Data Systems (SEEDS) formulation activity established a study to determine the opinion of the ESE community regarding the potential role of reuse in the development of future ESE data systems. The study included three steps. First, a range of options was determined through discussions with community practitioners and industry experts. These options included “status quo”, “improved clone and own”, “open source”, “service encapsulation”, and “product line” approaches to reuse. Second, community opinion regarding these options was solicited through workshops, surveys, and interviews. Finally, a process to enable reuse was characterized through community workshops.

The expert opinion of stakeholders in the ESE community and an assessment of the potential benefits and costs indicate that NASA’s ESE should initiate an effort to facilitate software reuse through “improved clone-and-own” and “open source” approaches. Because of divergence in community opinion, different approaches should be used in different environments: those in mission-critical environments strongly favor the improved clone-and-own approach, while those in mission-success (science and application) environments favor open source and service encapsulation approaches. Community opinion was strongly against attempting a product line approach—in spite of the potential for higher levels of reuse—apparently because past attempts have resulted in serious problems related to cost and responsiveness, and because the requisite organizational and funding structures would be difficult or impossible to implement across the diverse ESE community. The study team recommends against emphasizing a service encapsulation approach at this time—in spite of the potential to reuse operations infrastructure in addition to software assets—because the necessary Internet service protocols are not sufficiently mature and because of the somewhat limited applicability of this approach.

As part of this reuse initiative, the community indicated that the ESE should develop a coarse-grained, notional reference architecture with concrete details in a limited set of functional areas. The purpose of the reference architecture is to facilitate communications between component suppliers and potential users by providing common terminology and definitions for the subsystems that comprise an ESE data system. The community recommended against developing a fine-grained, specific reference architecture in spite of its theoretical benefits to component-level reuse and software interoperability because it would be too costly to develop and too constraining for use by the science community. As an aside, we note that the opinions of the ESE community

regarding reference architecture alternatives were not as strong as they were regarding reuse alternatives.

The study team performed a cost savings sensitivity analysis, which indicated that a significant return on investment is possible if a reuse initiative could make even modest improvements in how often and how much reuse is employed in future system development efforts.

The next step is to define and initiate processes to implement the community recommendations. These should be community-owned, non-prescriptive, scalable, practical processes that start simply and evolve, emphasize directly enabling reuse over infrastructure activities, and rely on competition and peer review rather than mandates to drive reuse appropriately. In keeping with these principles, SEEDS should competitively select and fund community reuse implementation projects and establish incentives that overcome artificial barriers to reuse.

To help kick off and further define these processes, the ESE should establish two reuse working groups: one focused on the improved clone-and-own approach in mission-critical environments, and one focused on the open source approach in mission-success environments. These working groups could be responsible for recommending specific reuse initiatives, and for working in the areas of outreach and education, support/enablement, and policy change to further enable reuse. The working groups should specifically be responsible for development of the recommended reference architecture as part of their support/enablement activities.

A separate body such as a SEEDS Integration Office could be responsible for prioritizing and approving reuse initiatives, for selecting and guiding community reuse projects, and for administering reuse incentives. It may also conduct some reuse outreach and education activities. The Integration Office could include a small technical team to support all reuse-related activities. Again, community input should be solicited to further define and continuously evolve the responsibilities of the Integration Office and the working groups.

Contents

Preface

Purpose and Scope

This document summarizes the activities and results of the SEEDS Reuse and Reference Architecture Study conducted as part of the SEEDS formulation effort.

Organization of This Document

The study results have been summarized in three levels of detail to suit the needs of various readers:

- The Executive Summary provides a complete, concise summary of the study results;
- Sections 1, 2, and 3 provide a more comprehensive summary of the study. Section 1 "1 Background" provides some useful context, including definitions of terms as used in the study. Section 2 "2 Current and Recommended Approaches" summarizes current activities related to the study, and (most importantly) the ESE community opinion on alternative approaches to reuse and reference architecture. Sections 2.3 "2.3 Expected Benefits" and 2.4 "2.4 Investment Costs and Organizational Fit" discuss the recommended approaches in terms of each of the five main evaluation criteria used in the study. Section 3 "3 Process & Next Steps" attempts to derive a straw process from the community recommendations as the starting point for further work.
- Sections 4 through 8 serve as an appendix, and contain additional detail on various aspects of the study.

Revision History

Version	Date	Description of Change
1.0	08/12/02	Initial draft for Formulation Team review.

Contributors

Name	Role
Gail McConaughy/GSFC	Study team lead.
Mark Nestler/GST	Study support lead.
David Isaac/BPS	Study support.
Nadine Alameh/GST	Study support.
SEEDS Public Workshop Attendees	Study input and recommendations. (See workshop attendee list for individual names.)

1 Background

The following sections provide some background on the study, including the motivation, reservations, definitions, goals, and approach.[2]

1.1 Motivation

Achieving the ESE science and application goals[3] over the next ten years requires NASA to provide new capabilities to the ESE community. We must provide data systems for additional missions that will satisfy new Earth science data needs. We must simplify the process of fusing diverse data for use in increasingly sophisticated models[4], interdisciplinary science, and new applications. And we must make access to Earth science data easier and faster to meet the unique needs of new applications[5].

Of course there are numerous challenges. Flat budgets and ongoing missions leave little room for new systems development. Systems large enough to handle Earth science data traditionally are expensive to build, take years to deliver, and lack the flexibility desired by scientists.[6] And the trend toward smaller, distinct missions has the potential to lead to redundant, stove-piped system development efforts.

Fortunately, there are also some exciting opportunities. Community expertise that has been underutilized in the past can be tapped to develop systems more effectively and efficiently than before. Information technology advancements not only have the obvious immediate cost benefits, but also make it feasible to put substantial storage, processing, and communications capability directly into the hands of end users and thereby remove certain organizational and logistical impediments to system development and usage. The successful deployment of numerous instruments allows us to shift our focus from simply getting data to exploiting it. Numerous mission systems provide a wealth of software assets that can potentially be reused to meet the needs of future missions. And last but certainly not least, interoperable Internet technologies have greatly improved our ability to create and utilize distributed data holdings and associated services.

[2] Additional background can be found in the study document “NewDISS Reuse & Reference Architecture Study: Analysis Approach”.

[3] The top-level science goal is to “observe, understand, and model the Earth to understand how it is changing and the consequences for life on Earth”. The top level application goal is to “expand and accelerate the realization of economic and societal benefits from Earth science, information, and technology.” See “Exploring our Planet: Earth Science Enterprise Strategic Plan”, Jan 2002.

[4] For example, the combination of wind and sea surface temperature to improve weather models.

[5] The tolerance for latency of data delivery in many applications can be lower than science uses. Weather-related applications are a prime example where use of near real time data is common.

[6] Some of this is inherent in Earth science and applications: non-routine analysis on large data sets is hard to do. But some of it may be due to our approach in the past to building these systems.

At the intersection of these goals, challenges, and opportunities is the need (and ability) to “implement an open and distributed information system architecture that will include Principal Investigator processing in the mix of science data processing providers, and tie together diverse creators and users of higher level information products.”[7] This architecture should both enable and benefit from increased, accountable science community participation; flexibility and responsiveness relative to science needs; and effective use of available resources.[8] We believe that an ESE reference architecture is the means of capturing and communicating this vision, and that software reuse is key to realizing the vision within practical budget constraints.

1.2 Reservations

Although many organizations have realized significant cost savings and other benefits from software reuse, it was not obvious to the study team that these same benefits would be realized within the ESE. Many software reuse approaches are dependent on organizational structures and funding approaches that simply are not practical across the diverse and nationally distributed organizations that comprise the ESE community. Specific concerns included the cost of developing reusable software (and who would bear the cost); barriers to reuse arising from quality and schedule risks; potential negative effects on innovation and community participation; and the possible rapid decay in reusable asset value due to technology changes.

Similarly, although the case for defining a reference architecture in general appears strong, we did not embark on the study with the foregone conclusion that it is the right thing for the ESE. Our concerns included the cost and time needed to develop the reference architecture; questions about its actual utility once developed; potential negative effects on flexibility, innovation, and community participation; and the potential to hinder—or be made irrelevant by—technology infusion. Further, the term “reference architecture” is so loosely defined that even an agreement that one should be developed says nothing about what form it should take. We designed the study to address all of these issues.

1.3 Definitions

During the course of the study it became apparent that community opinion on reuse and reference architectures depended largely on how one defined these terms and the approach taken to realize the associated goals. To address this, we defined reuse and reference architectures within the context of the study and offered a variety of approaches for stakeholders to evaluate. We also suspected and confirmed that opinions within the community differed depending on the work environment. We provide definitions of the

[7] Exploring our Planet: Earth Science Enterprise Strategic Plan”, Jan 2002.

[8] Details of the linkage between the study recommendations and these expected benefits are contained throughout this document. For example, we note that architectures are critical to enabling software reuse (which in turn provides more “effective use of available resources”), and that reusable software can greatly speed adoption of an architecture.

two most important community segments, “mission-critical” and “mission-success”, which are referred to throughout this document.

1.3.1 Reuse

Reuse is the act of taking a functional capability used in (or provided by) one system or mission and employing it in another system or mission. This broad definition is intended to encompass a variety of techniques that have the potential to reduce future data system costs, not simply libraries of reusable software components. For example, employing an entire existing system (including software, hardware, and operational processes) to support a new mission would fit this definition of reuse.

For the study, we considered four approaches to software reuse.[9] We define a *Clone & Own* approach as copying code and associated artifacts for use in another system, where they may be independently modified and maintained. We recognize two variants of this approach: *Ad Hoc Clone & Own,* in which development teams employ a clone & own approach using their personal knowledge of available systems, and *Improved Clone & Own,* in which processes and mechanisms are put in place to facilitate Clone & Own practices. An *Open Source* approach is similar to Clone & Own, but development is typically distributed across multiple organizations and a person or organization is assigned to maintaining a consolidated repository into which additions or fixes are integrated. A *Service Encapsulation* approach entails wrapping a complete, operational capability with a network-accessible interface so that the capability can be used in-place (rather than re-implemented). A *Product Line* approach is based on reusing a set of core software components intentionally designed for a family of systems; the components are modified and maintained only by the organization responsible for the core components.

	Clone & Own	Open Source	Service Encapsulation	Product Line
Distribution mechanism	Source	Source	Network Interface	Binary
Code modified in end system	Allowed	Allowed	Disallowed	Disallowed
Integrated asset repository	No	Yes	Yes	Yes
Designed for reuse	Rarely	Usually	Varies	Always

Figure 1.3-1 Characteristics of different software reuse approaches considered in this study.

1.3.2 Reference Architecture

Architecture is defined as the structure of components, their interrelationships, and the principle guidelines governing their design and evolution over time (i.e., the components, connections, and constraints of a system). A *reference architecture* is a generic architecture that provides coherent design principles for use in a particular domain (in this case, Earth science). It is used as a reference (for either guidance or compliance purposes) when developing an architecture for a specific system.

[9] Detailed definitions can be found in “5.2 Reuse Alternatives” in the Appendix.

Reference architectures can have varying levels of specificity and granularity. We define a *notional* architecture as the least specific. It identifies components and allocates functional requirements to them, identifies the most important connections, and may identify a few constraints. We define the next level as a *concrete* architecture. It defines the actual services (and their parameters) for each component, all major connections, and some constraints. The last level we define as a *specific* architecture. It goes beyond a concrete architecture by defining the invocation mechanism for each service and establishes constraints (such as formal standards for interfaces) to ensure software component interoperability. We also define *coarse, medium,* and *fine* levels of granularity, corresponding roughly to a subsystem, functional component, or software module level of detail in the architecture.

	Notional	Concrete	Specific
Component definitions	Descriptive	Allocated Requirements	Functional Specification
Component services	Undefined	Named	Specification
Data flows	Descriptive	Comprehensive	Format Standards
Communications infrastructure	Undefined	Protocol Standards	Ancillary Service Specifications[10]
Physical components/systems	Undefined	Types Defined	Instances Defined

Figure 1.3-2 Characteristics of different reference architecture approaches considered in this study.

1.3.3 Community Segments

As expected, responses collected by the study team from the ESE community during the public workshops and one-on-one interviews show a clear diversity in the community opinion. The strongest opinion differences fell along the lines of the following identified community environments:

- *Mission-critical* environments are driven by launch schedules and a need for daily, highly reliable production or archiving needs. Examples include SIPS and DAACs for standard products and high volume distribution.
- *Mission-success* environments are driven more by need for research and innovation in science, applications, or information systems; the need to experiment with differing products, approaches, mechanisms; and the need to adapt to new understandings. Examples include ESIP-2s, -3s, analysis, etc.

Classifying the stakeholders' opinions according to these two community environments ensures that the opinions collected reflect the needs, requirements and constraints of each community, hence ensuring that a one-size-fits-all recommendation is avoided. It is also important to note that although the presentation of the community opinion described in this report is based on this mission-critical versus mission-success segmentation, the study team understands that some community members participate strongly in both types

[10] Ancillary services include, for example, directory services for identifying available services.

of activities. For the purpose of aggregating community opinion for this study, stakeholders were asked to represent the one environment which they most self-identify with (often based on their primary funding sources).

1.4 Study Goals and Approach

This study seeks to determine if reuse and/or reference architectures can provide the following return on investment:

- Reduction of the cost of supporting future missions, science, and applications;
- Increased flexibility and responsiveness to new missions, science, and applications; and
- Increased effective, accountable community participation in system development and operations.

If the answer is "yes", what processes would best move ESE toward these goals? How can the community and NASA implement those processes?

The approach to this study centers around three key themes:

- Reliance on the opinion of experienced stakeholders in the ESE community, with particular emphasis on practical experience with actual missioncross-system and cross-project reuse;
- Examination of lessons-learned and recommendations from related activities and expert resources such as the Carnegie Mellon Software Engineering Institute and the OpenGIS Consortium; and
- Incorporation of feedback obtained from the ESE scientific community through interviews and workshops.

Initially, the study consisted of a preliminary trade study and analysis. The study team interviewed and surveyed selected ESE stakeholders and data system developers, and reviewed documented case studies, reports and papers. Groups consulted included system developers from SeaWifs and TSDIS, and software engineering groups such as the Carnegie Mellon Software Engineering Institute. Case studies included those from companies such as McDonald Detweiler and Motorola. Based on information gathered, the study team was then able to identify a range of software reuse and reference architecture options. Criteria to evaluate these options were also developed. The specific options and evaluation criteria are described in Section 2 and the Appendix.

After developing options and evaluation criteria, the study team then focused on formal solicitation and compilation of community views toward software reuse and reference architectures. These views were gathered via multiple workshops and further one-on-one interviews. Community members consulted included individuals from various ESIPs, Distributed Active Archive Centers, and ongoing and future ESE system development efforts. The overall community viewpoint was then published.

Subsequent activities will examine community-based processes for enabling reuse and defining a reference architecture. Consensus-based processes will be developed in detail by stakeholders in the ESE community. Multiple working groups of stakeholders will examine these processes with the assumption that "one size does *not* fit all" types of .environments. It is expected that this aspect of the working groups' activities will be evolutionary, but ultimately, these groups will provide NASA headquarters with final recommendations. Then, if headquarters approves and funding is supplied, the working groups will be charged with prioritizing and implementing the recommended community-based processes.

2 Current and Recommended Approaches

The study team examined activities related to software reuse and reference architectures to determine what approaches were currently being employed. We then collected opinions from the ESE community on whether or not any of the approaches would support the goals of SEEDS. Based on practical experiences with the best approaches identified by the ESE community, we attempted to qualify and quantify the benefits that could be expected. The information gathered and generated during each of these steps is summarized in the following sections. Additional details are provided in the appendix.

2.1 Current Activities

Our survey identified a variety of divisional, contract, and research activities at NASA related to software reuse. These include activities within the Flight Dynamics Division, the Software Engineering Laboratory, the Flight Software Division, various missions (TRMM, MLS, QuikTOMS, GPM), the ECS project, DODS/OPeNDAP, and various research projects and cooperative agreements. We also noted significant reuse efforts within the DoD and commercial companies, both largely focused on the product line approach. DoD activities include the Central Archive for Reusable Defense Software (CARDS) program, the Software Technology for Adaptable, Reliable Systems (STARS) program, the Software Reuse Initiative and associated Reuse Information Clearinghouse, the Portable Reusable Integrated Software Modules (PRISM) program, and the NRO's Control Channel Toolkit program. And of course there is a significant amount of activity in the open source community, including the Open Channel Foundation (which hosts the NASA COSMIC software collection) and the Computational Technologies Project[11] software repository associated with the National HPCC Software Exchange.

The results of the survey indicate that there is a significant opportunity to realize additional benefits from software reuse. Where reuse was employed at NASA, good results were often achieved in terms of cost savings and quality improvement. However, reuse was not employed as often as it could be. One reason for this is that support mechanisms for reuse are lacking, so the level of reuse for a particular data system depends largely on the individual relationships of developers working on different systems. The survey also highlighted that different approaches to reuse are suited to

[11] Formerly the Earth and Space Sciences Project.

different environments. In particular, most of the DoD and commercial initiatives focused on a product line approach, which has requirements for success (including up-front investment, non-mission-oriented organization and funding structures, and requirements stability) that are not obviously suited to the mission-oriented, organizationally distributed, innovative nature of the Earth Science Enterprise. By contrast, most of the successful ESE data system reuse efforts use a "clone & own" approach, which affords greater flexibility in tailoring components to meet new requirements. What is needed is a process that builds on approaches proven to be successful in the ESE environment so that more ESE data system development efforts can consistently realize the potential benefits of software reuse.

Our survey identified 18 activities in standards groups, ESE projects, and other government organizations that have developed or are working on reference architectures relevant to NASA's ESE. These include the OGC OpenGIS Service Architecture, the FGDC Standards Reference Model, the FGDC Geospatial Interoperability Reference Model, the NIMA USIGS Objective System Architecture Description (and related documents), the SDI Implementation Guide (the "cookbook"), the Global Grid initiative's Open Grid Services Architecture, the ISO Geospatial Framework & Reference Model (and related documents), the CCSDS Open Archive Information System, the Interoperability Clearinghouse, the Open-Source Project for a Network Data Access Protocol (OPeNDAP), the NASA Renaissance Ground Data Systems Architecture (now under the OMG SOTG), the NASA Earth System Modeling Framework, the DARPA Domain Specific Software Architecture program, the DARPA Intelligent Integration of Information reference architecture, the DISA Global Information Grid program, and the DOE Common Component Architecture.

Although many of these would likely be used as a reference by some developers of future ESE data systems, there is currently no comprehensive, consistent guidance or compliance requirements within NASA's ESE that would fulfill the role of a reference architecture as envisioned in the "Motivation" section above. What appears to be missing is not so much another architecture effort, but leveraging of existing efforts to create a reference architecture that is tailored to Earth science, that is more commonly used as a single point of reference within the ESE community, and that more directly provides tangible benefits such as cost reduction through software reuse.

2.2 Community Opinion

This section summarizes the ESE community opinions regarding (1) the choice and suitability of the identified reuse and reference architecture alternatives (collected during Phase I of this study), and (2) the key process recommendations (collected during Phase II of this study).

2.2.1 Community Opinion on Reuse and Reference Architecture Alternatives

In Phase I, the study team solicited the opinion of eighteen stakeholders in the ESE community (with good representation from DAACS, SIPS, ESIP-2s and ESIP-3s) to evaluate the identified alternative approaches to software reuse and reference architecture development (summarized in Section 1.3 and detailed in the Appendix)). For every alternative approach, the stakeholders were asked to rate as positive, neutral or negative each of the following equally-weighted evaluation criteria[12]:

1) Potential for cost savings;

2) Potential to improve system flexibility and responsiveness to science and application needs;

3) Potential to promote increased, accountable community participation;

4a) Suitability for use within the ESE mission-critical environment;

4b) Suitability for use within the ESE mission-success environment; and

5) Investment cost.

The following sections summarize the results of the evaluation. We first present the aggregate opinion of the mission-critical and mission-success communities about the proposed alternatives, taking into account all the evaluation criteria listed above. Next we show the aggregate opinion of each community, focusing solely on the suitability of the identified options to each environment. Such an approach highlights the differences of opinions between the two communities and the necessity of having each community define an approach that is tailored to its own needs.

The sections below show that both communities are not satisfied with Status Quo and they both agree that something needs to be done. Interestingly, the community opinion regarding Reference Architecture alternatives were not as strong as they were for reuse. Furthermore, there was a clear divergence of community-desired approaches, implying the need for different approaches for the two identified environments. Additional details including a summary of the written comments as well as the survey itself can be found in the Appendix.

Reuse

The common opinion of the stakeholders and development teams surveyed was simply “software reuse should be done.” One stakeholder commented that it is a waste of money to re-invent every mission data system. Indeed, Figure 2.2.1-1 shows that both the mission-critical and the mission-success communities are not satisfied with Status Quo.

Figure 2.2.1-1 also shows that the mission-critical community is strongly in support of an Improved Clone & Own approach. Other reuse alternatives were viewed as having fewer

[12] For the purpose of analytically representing and aggregating these ratings, the study team assigned the respective values of 1, 0 and -1 to the positive, neutral and negative responses of stakeholders. The figures shown in this section are based on the sum of these values for each alternative approach (aggregated by community and/or by evaluation criteria).

benefits or more serious problems. While the Open Source option elicited some support, there were concerns that the lack of control in an open source environment would pose difficulties in estimating system costs and schedules. The evaluation survey also shows that the Product Lines approach was not rated well by the mission-critical community in large part because some stakeholders associated it with large, centralized contract development approaches that are perceived as costly and inflexible. This negative rating is interesting in light of the fact that this approach receives perhaps the most attention within DOD and commercial software reuse initiatives, and that the evaluators were told that this is the approach recommended by the Carnegie Mellon Software Engineering Institute.

As for the mission-success community, Figure 2.2.1-1 indicates that stakeholders in this community are almost equally in favor of the Service Encapsulation and the Open Source options. As for its opinion about Product Lines, the mission-success community seems to be in disagreement on that option (reflected by a zero value on the chart, which the study team confirmed is the result of opposite-value responses canceling each other out).

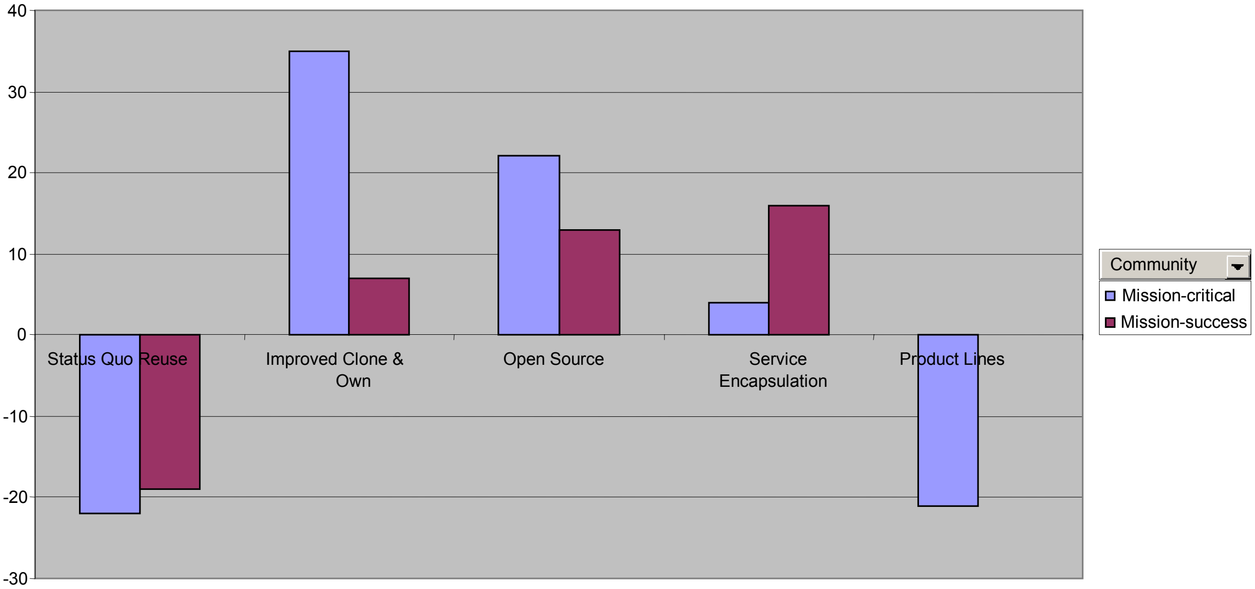


Figure 2.2.1-1 Mission-critical and mission-success community opinions regarding alternative reuse approaches, taking into account all evaluation criteria.

Figure 2.2.1-2 and 2.2.1-3 allow us to gain a better understanding of the opinion of each community regarding the suitability of each Reuse option to both environments. These figures confirm that the options preferred by each community do often differ from the ones proposed to it by outside communities. For instance, while the mission-critical community seems to be strongly in favor of the Improved Clone & Own approach for itself, the mission-success community considers the Product Lines approach more suitable for that environment (Figure 2.2.1-2). On the other hand, while the mission-success community seems to be equally in favor for the Service Encapsulation and the

Open Source options for itself, the mission-critical community considers the Improved Clone & Own option more suitable for that environment (Figure 2.2.1-3).

It is particularly interesting to note how Improved Clone & Own was the mission-critical community's most suitable Reuse option for both environments. This observation suggests that the surveyed stakeholders' responses often reflected their experiences within their own environments even when considering options for outside environments. This in turn emphasizes the need for this self-for-self and cross-opinion presentation of the survey responses.

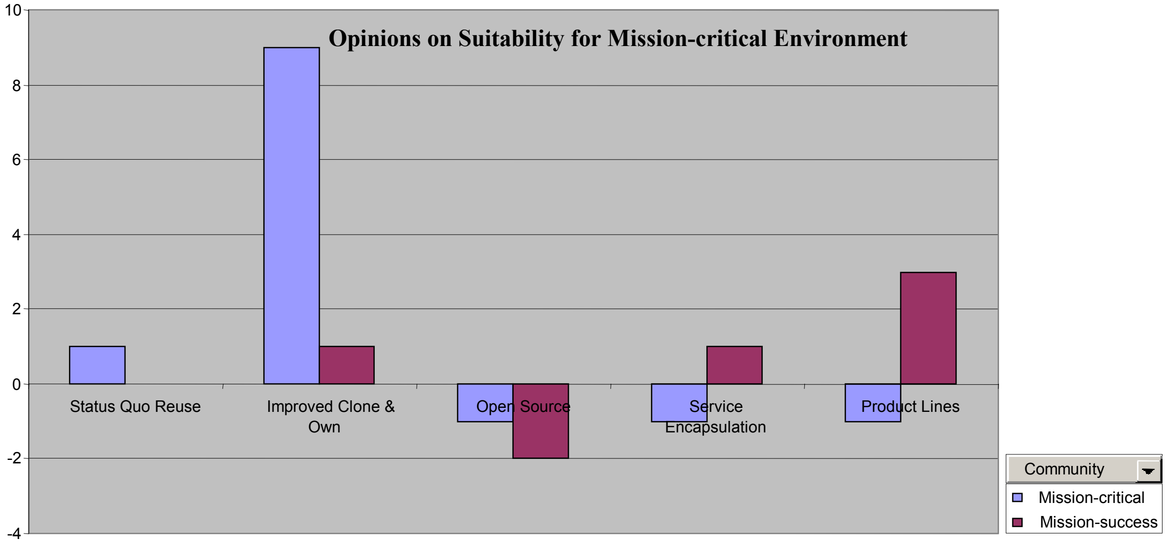


Figure 2.2.1-2 Mission-critical and mission-success community opinion regarding the suitability of each reuse option to the mission-critical environment. Note that the preferred approach of the mission-critical community for itself (Improved Clone & Own) is different from the one suggested to it by the mission-success community (Product Lines).

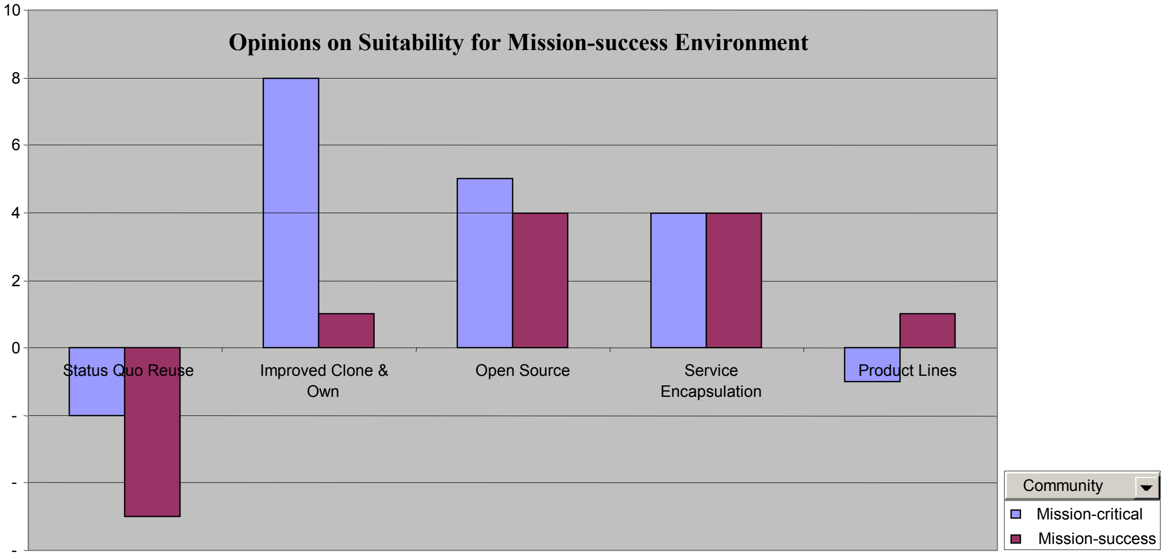


Figure 2.2.1-3 Mission-critical and mission-success community opinions regarding the suitability of each reuse option to the mission-success environment. Note that the preferred approaches by the mission-success community for itself (Open Source and Service Encapsulation) are different from the one suggested to it by the mission-critical community (Improved Clone & Own)).

Reference Architecture

As for the community opinions regarding Reference Architecture options, these were not as strong as those gathered for the Reuse options. Nonetheless, the survey indicated that the *status quo* is probably not satisfactory as far as a Reference Architecture is concerned (especially for the mission-success community). Moreover, a notional or concrete Reference Architecture would help meet ESE goals as long as it is designed to drill down to more detail in selected functional areas. The survey also indicated that a Reference Architecture should be coarse grained, dealing with whole subsystems and their relationships rather than, say, being concerned with the way individual software modules are connected. Fine-grained architectures were not rated well largely because of concerns that they would be inflexible and thereby inhibit innovation, community participation, and technology infusion. The low rating of fine-grained architectures also emphasizes the community's interest in keeping the architecture at a high level of detail.

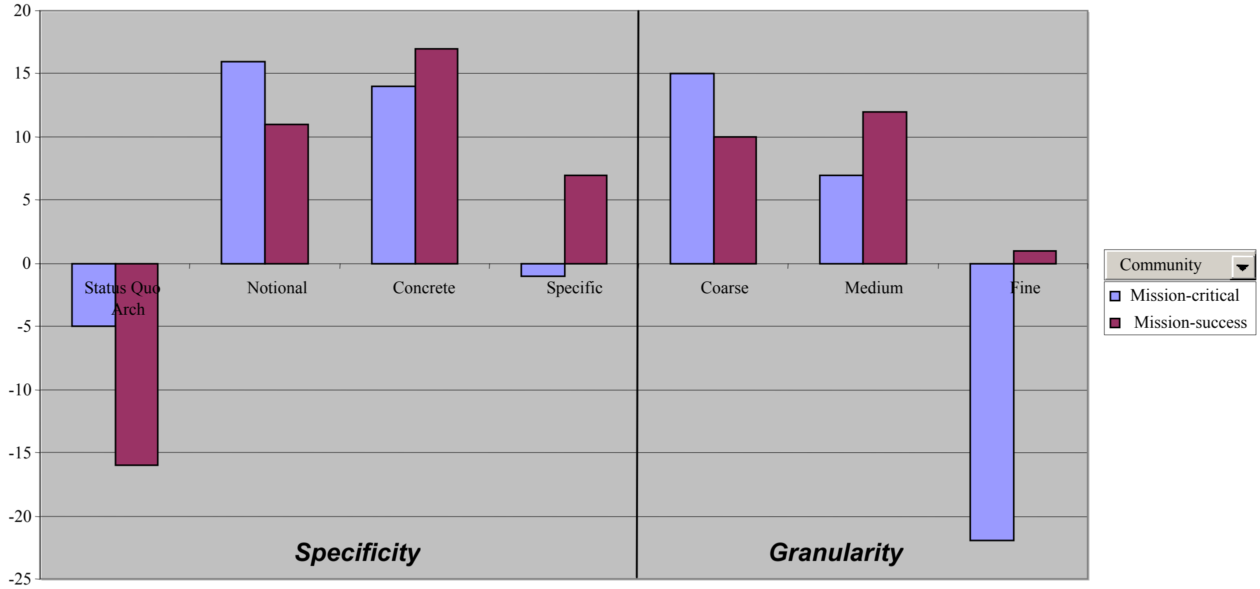


Figure 2.2.2-1 Community opinion regarding different levels of specificity and granularity for a Reference Architecture. Most indicated that a notional or concrete reference architecture would help meet ESE goals, but that the *status quo* (no reference architecture) or a very specific reference architecture would not. Most indicated that the architecture should be defined only at the coarsest (i.e., subsystem) level, with some "drill-down" to a medium (functional component) level where appropriate.

Figures 2.2.2-2 and 2.2.2.-3 confirm that the differences in opinions between the mission-critical and the mission-success communities were less pronounced than for the Reuse alternatives. Both communities find Concrete and Specific architectures suitable for the mission-critical environment, provided additional detail is added only as needed in selected functional areas. As for the suitability for the mission-success community, the mission-success community is against continuing with status quo, favoring a Notional or Concrete architecture for itself. Finally, unlike the mission-critical community, the mission-success community seems to be in disagreement about a Fine architecture for itself.

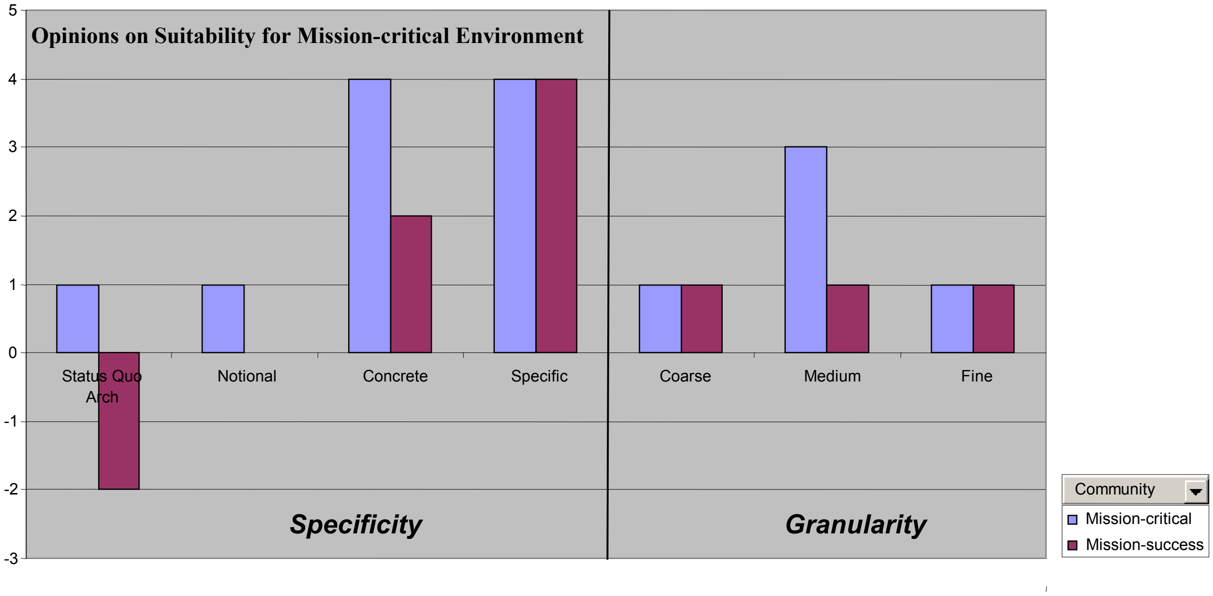


Figure 2.2.2-2 Mission-critical and mission-success community opinion regarding the suitability of each Reference Architecture option to the mission-critical environment.

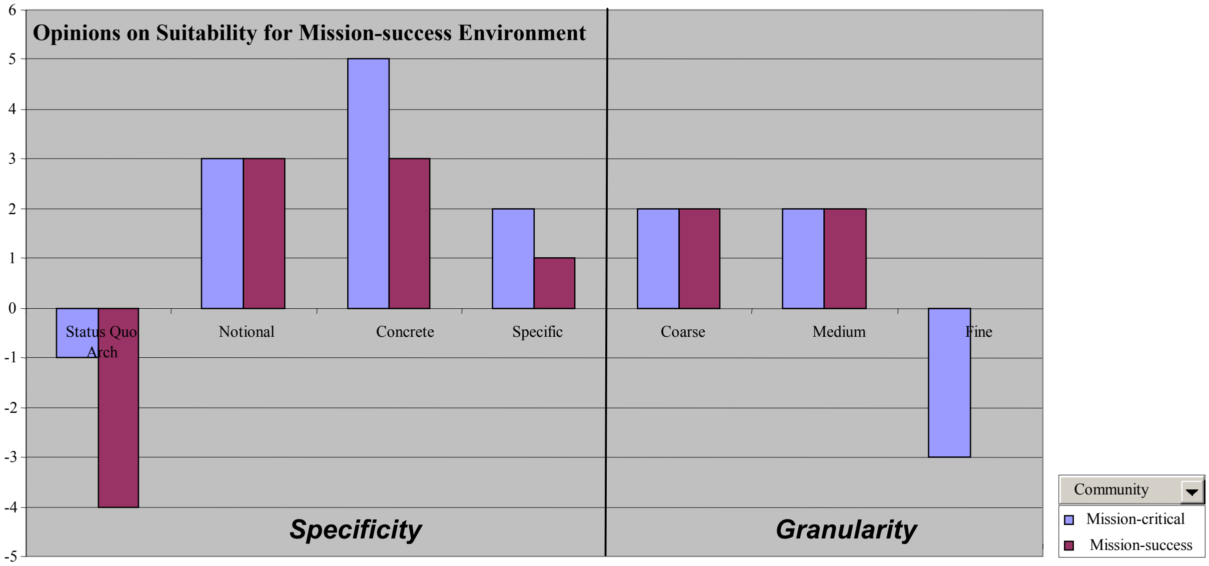


Figure 2.2.2-3 Mission-critical and mission-success community opinion regarding the suitability of each Reference Architecture option to the mission-success environment.

2.2.2 Community Opinion on Key Process Characteristics

The study team solicited the opinions of stakeholders regarding the reuse process definition in the areas of guiding principles, contributing factors, program and technical strategies, evolutionary approaches and reference architecture use. The community opinion highlights captured by the study team on key process recommendations include the following:

- Do something, because reuse across projects rarely happens by itself. In many cases, only a small amount of additional effort or funding may be needed to make valuable software available outside the group that developed it. In other cases, significant barriers (especially intellectual property policies) may have to be removed or circumvented.
- Start with a simple process and engage all stakeholders in refining and evolving it. Leverage existing resources, infrastructure, forums, etc. as well as lessons-learned from similar initiatives to ensure that real results are achieved quickly and cost effectively.
- Use competition and peer-review rather than blanket policies to drive reuse to help ensure that reuse always serves the ESE goals and does not become an end in itself.
- A record of authorship and access to authors is essential for an asset to be reused. In this regard, code is treated much like science data: usage often boils down to quality and trust in the source.

More detail about these and other recommendations can be found in the Appendix.

2.3 Expected Benefits

The survey indicates that software reuse—employing an improved clone & own or open source approach and facilitated by a coarse, notional (or concrete) reference architecture—should provide benefits for future ESE data systems including cost savings; increased flexibility and responsiveness to science and application needs; and increased, accountable community participation. The following sections provide some quantification of, and perspectives on, these expected benefits.[13]

2.3.1 Potential Cost Savings

To reduce software development costs, there are generally two basic choices: improve productivity in order to develop the same amount of software faster, or reduce the amount of software developed. Software reuse directly addresses the latter option, with the potential for substantial cost reductions that are directly proportional to the amount of software reused. In an era of constrained budgets, such savings are essential to meeting science requirements, providing new capabilities, and supporting technology infusion.

[13] Note that these are the benefits that should result from the recommended approaches, not the theoretical benefits of software reuse in general. As such, these expected benefits should be used as measures of success for any subsequent ESE software reuse initiative that is based on these recommendations.

An ESE reuse initiative could potentially provide cost savings in two primary ways: by increasing the percentage of development efforts employing reuse, and by increasing the amount of reuse within each development effort. Other factors play a significant role, but are harder to influence with a reuse initiative. The study team performed a cost savings sensitivity analysis using the Poulin/Caruso model[14] By gradually increasing the percentage of reuse over eight missions from an estimated current value of 30% to an achievable level of 60%, and by ensuring that all those missions employ reuse (rather than only half as we would expect today), the ESE could free up at least an additional 25% of the total custom software development costs for other uses.[15]

An ESE reference architecture could provide cost savings in three ways: reuse of assets, increased competition, and improved development efficiency.

A reference architecture both enables software reuse and is itself reuse. If two systems are based on the same reference architecture, there is a better chance that a software component developed for one could be used in another.[16] Any increase in reuse has a dramatic effect on cost savings, since it typically costs 80% less to reuse a compatible component than to develop it from scratch.[17] In a "clone-and-own" approach to reuse, a reference architecture is arguably less important than in a product line approach, because one simply adopts the architecture of the cloned system. Still, a reference architecture provides the basis for gradually converging different cloned systems toward a compatible architecture, increasing the opportunity to reuse components not only within one cloned system family but also across families. Of course, when a reference architecture is used to develop a system, this is itself a form of reuse. And because the design phase can be a substantial portion of the overall system development cost, the cost savings relative to developing an architecture from scratch should still be substantial.

A reference architecture helps increase competition by providing the technical basis for interchangeable (competing) components. The recommended notional or concrete reference architecture, however, will not have sufficient specificity to enable plug-and-play functional components, so there will likely be little increase in competition at that level. However, there may be some benefits to competition at the service provider level because of improved compatibility among subsystems and better "apples-to-apples" comparisons of provider capabilities (using the services in the reference architecture as a checklist).

[14] Poulin, Jeffrey S. Measuring Software Reuse: Principles, Practices, and Economic Models. Addison Wesley. 1997.

[15] Such improvements should be possible by making components more widely available, helping to improve component documentation, sharing successful approaches and tools, etc.

[16] See http://www.hpl.hp.com/techreports/98/HPL-98-132.pdf and http://www.sei.cmu.edu/pub/documents/96.reports/pdf/tr018.96.pdf for discussions of the linkage between architecture and reuse.

[17] Tracz, W., "Software Reuse Myths," ACM SIGSOFT Software Engineering Notes, 13(1), 1998, pp. 17-21.

A reference architecture can help improve development efficiency in a variety of ways, but the most important is by providing the technical basis for cleanly partitioning the system development effort among smaller teams of experts. The resulting cost savings can be significant: software development productivity on a smaller subsystem (say 100 KLOC[18]) can be 25% to 30% higher than on a system just four times larger[19]. The opportunity to use in-place teams rather incurring the cost of ramping up a whole new organization is an added bonus.

2.3.2 Flexibility and Responsiveness

Software reuse can improve flexibility and responsiveness in two ways: by reducing system development time, and by leveraging a broader community to develop needed enhancements.

First, software reuse improves responsiveness by reducing system development time. In fact, reduced "time to market" is often the primary driver for software reuse initiatives at commercial firms. Assembling a system (at least partially) from existing components not only has the obvious benefit of reducing the amount of software that has to be developed, but can also have the additional benefit of providing a solid starting point to jump-start a development project. The ability to deliver ESE data systems more quickly would undoubtedly be of interest to not only the science community, but also to program managers anxious to get data system development off the critical path.

Second, software reuse can help engage and leverage a broader community to develop needed enhancements. For example, the open source concept allows anyone in the ESE community to respond to specific needs and offer solutions to the source repository. And even with the improved clone and own approach, software "reusers" can offer enhancements back to the software authors or to other "reusers". Each project is free to adopt the enhancements or not depending on their needs and constraints.

A reference architecture can improve flexibility and responsiveness in three ways: by improving responsiveness to new requirements, by improving support for the requirements of multidisciplinary science and applications, and by enabling technology infusion.

A reference architecture can improve responsiveness to new requirements by cleanly partitioning the functionality of an ESE data system so that development can be performed by smaller, autonomous, expert teams. Such teams generally are faster at interpreting user requirements and implementing appropriate solutions within their specialty area. To realize this benefit, the reference architecture would need to be used as the basis of partitioning ESE data system development contracts. For example, the scope of a single contract might be for a data server, rather than a complete end-to-end system.

[18] KLOC = Thousand Lines of Code

[19] S. McConnell, *Rapid Development: Taming Wild Software Schedules*, pp. 194-196.

A reference architecture can improve support for the requirements of multidisciplinary science by improving interoperability. Although data interoperability (achieved through data format standards, a subset of a technical architecture) has received significant emphasis in the past, a more comprehensive architecture is needed to provide interoperable services. This will become much more important as data systems continue to provide higher levels of service (e.g., subsetting).

A reference architecture can enable technology infusion by reducing compatibility problems and providing a stable foundation for the development of innovative capabilities. With fragmented and incompatible architectures, technology developed for one architecture is almost by definition incompatible with another, so the benefits of the technology will be limited to a subset of the ESE community. Further, the lack of a clear target architecture and risk of acceptance may put off technology developers entirely. To the extent that a reference architecture can define and eliminate reworking the mundane aspects of ESE data systems, the more real innovation should occur. While a very specific architecture could prove too constraining and thus inhibit technology infusion, the recommended notional or concrete architecture should cause fewer problems in this respect.

2.3.3 Increased, Accountable Community Participation

Software reuse should provide increased, accountable community participation in two ways: by providing processes for community members to contribute software components, and by enhancing the ability of community members to leverage existing assets to create their own data systems.

A software reuse initiative provides processes for community members to participate in a small way by providing a needed (reusable) component or enhancements to an existing component. The open source approach by definition encourages such increased participation. While this is possible today, the lack of a reuse process including proper incentives and the support of experts from the community makes such contributions unfavorable both in terms of potential benefits and risks. Participation will increase as more groups contribute components for reuse, especially when supported and encouraged by the reuse initiative.

A software reuse initiative also helps community members participate in a large way by making reusable assets available. This should help small teams participate by providing essential components for a data system that the team could expand upon to provide additional valuable data products or data analysis services.

A reference architecture should provide increased, accountable community participation in three ways: by enabling smaller teams to participate in the development process, by enabling science teams to offer data services, and by establishing norms for component capabilities.

A reference architecture enables smaller teams to participate in the development process by partitioning large ESE data systems into manageable chunks. Without this partitioning, only a limited number of large system integration contractors can credibly bid to build an end-to-end data system. In addition, it is well known that hidden, poorly documented, and proprietary interfaces between software components limits opportunities for different development groups to offer component solutions;[20] a reference architecture helps to ensure that critical interfaces are publicly defined.

A reference architecture also enables distributed teams to offer data services by providing an common definition of the components and to some extent (for the recommended approach) the manner in which components offer these services. Consider by contrast forty data systems with forty different architectures: if an interdisciplinary science or applications team wants to offer a unique data product, what should the interface to the new data services look like? A reference architecture reduces the need to develop numerous interfaces or a forty-first design, either of which might involve enough effort that some potential data providers simply would not participate.

Finally, a reference architecture increases accountability by providing community definitions and norms for component capabilities, which can be used as a measure of success for any component or service provider.

2.4 Investment Costs and Organizational Fit

While all the approaches to reuse and reference architectures could theoretically provide most of the benefits above, realizing those benefits in practice is another matter. In particular, it is important to note that the success of an approach within the limited scope of a single commercial organization says little about the likely success of that approach in something as broad and organizationally diverse as the Earth science community. As detailed in the following sections, the conclusion of the ESE community was that the Improved Clone & Own approach is best for mission-critical environments, and the Open Source approach is best for mission-success environments.

2.4.1 Suitability for Mission-Critical and Mission-Success Environments

Community opinion indicates that the Improved Clone and Own approach to software reuse is best for mission-critical environments. It provides proven software with known functional and performance characteristics, and less risk of latent defects. It also provides complete control over the code, allowing a project to make needed changes while isolating the code from unexpected or unwanted changes by others. The result is more certainty in development schedules and less risk of operational problems.

The Open Source approach to software reuse is most suitable for mission-success environments. It provides the opportunity for any organization to add innovations to base

[20] This is the basis of numerous lawsuits against IBM in the 1970s and 1980s, and Microsoft in the 1990s.

capabilities, and (optionally) to share these with others in the community. For those willing to trade version stability for capability enhancements, it also offers the opportunity to deploy enhancements contributed by others in the community.

A coarse, notional architecture is suitable to both mission-critical and mission-success environments because it provides the benefits described above while imposing few constraints. We expect that mission-critical environments might gravitate more quickly toward a concrete architecture, which will provide more software component interoperability at the cost of the effort and consensus building needed to get to the additional level of detail.

2.4.2 Investment Costs

The recommended approaches can be characterized as "light touch", and consequently the effort needed to implement them are small relative to both the other approaches and to the benefits provided.[21]

3 Process & Next Steps

3.1 Process Characteristics

A substantial amount of guidance from the community relative to a reuse process has been captured in Section 0 "7 Reuse Process and Next Steps". To summarize, input from the community indicated the following:

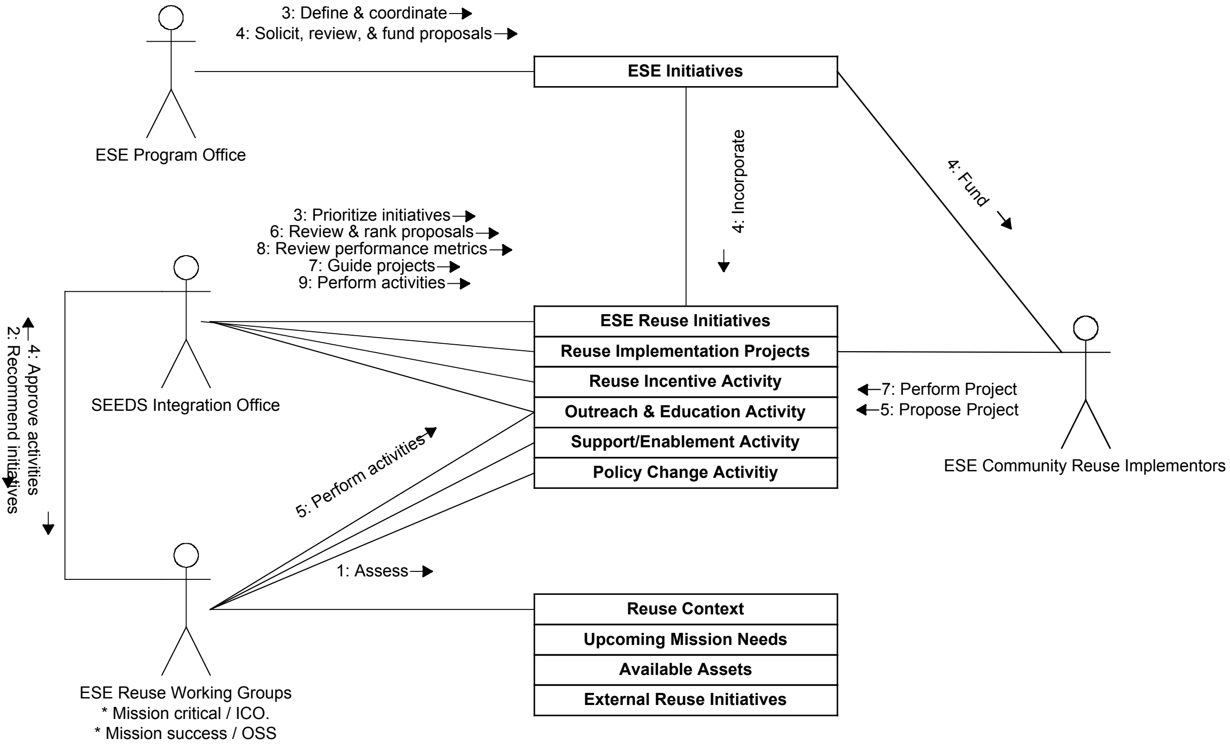
- **Principles.** A SEEDS reuse process should be a community-owned, non-prescriptive, scalable, practical process that starts simply and evolves, emphasizes directly enabling reuse over infrastructure activities, and relies on competition and peer review rather than mandates to drive reuse appropriately.
- **Contributing Factors.** There are currently a number of barriers to reuse, including project funding constraints, licensing issues, support concerns, security concerns, cultural issues, and communication issues. A process that removes some of these barriers could significantly improve the level of reuse within the ESE.
- **Program Strategies.** A reuse process should emphasize community-based working groups, fund actual reuse activities rather than infrastructure activities, and establish incentives to encourage reuse. To a lesser extent, a reuse process should facilitate sharing of knowledge related to reuse and reusable assets, provide some institutional support, and establish/revise policies to further enable reuse. Again, the principle of starting with a small, simple process and building on what works was emphasized.

[21] The SEL estimates that their object-oriented development initiative costs were 5% of the system development budget. While this amount is small relative to the benefits claimed, a reuse initiative using the recommended approaches would be far less sweeping and should cost even less.

- **Technical Strategies.** A reuse process should employ a variety of technical strategies including information sharing (through workshops, contact directories, published success stories and best practices, checklists, etc.), quality indicators (e.g., identifying component authors), and direct funding (e.g., of documentation/generalization/support for components used across projects). At the level of technical strategies, differences between environments become more apparent. For example, those considering the improved clone and own approach for mission-critical environments favored in-place support from authoring organizations and no component library, while those considering the open source approach for mission-success environments favored establishing an open source infrastructure utilizing existing tools. The input indicates that technical strategies focused on methodology, policy enforcement, and automatic programming are not appropriate. The community emphasized that, regardless of the technical strategies employed, it is important to focus on components with a high likelihood of reuse.
- **Evolution.** The process should evolve primarily in terms of the definition of the process itself (starting simply and learning from experience), and also in terms of focus (i.e., which functional areas have the highest potential payback).
- **Reference Architecture Use.** It seems clear that the community is interested in knowing what components are available to meet a specific need, and that a reference architecture should, above all else, provide the definitions needed to ensure effective communications between component suppliers and component users.

3.2 Notional Reuse Process

The following diagram depicts how specific reuse initiatives could be identified and pursued over time. It is important to emphasize that this notional process is based on the community input summarized above, but has not itself been subject to community review. The key elements of this process are a set of small ESE Reuse Initiatives that are implemented through a variety of reuse projects and activities; ESE Community Reuse Implementers who actually perform reuse implementation projects; a SEEDS Integration Office that (among its other duties) is responsible for the reuse initiatives; and community-based ESE Reuse Working Groups that perform certain reuse activities. .}



Additional discussion of this notional process can be found in Section 0 “7.2 Notional Reuse Process”.

Appendix: Study Details

4 Current/Related Activities

The study team identified current activities related to the study topic to determine if those activities were sufficient to meet the needs of SEEDS, and to provide ideas for any activities initiated as part of SEEDS (if needed). The following sections summarize current activities related to reuse and reference architecture.

4.1 Reuse

One component of this study involved identifying recent software reuse activities. A broad cross-section of activities is desirable, so we examined not only NASA Earth Science Enterprise activities, but also defense and commercial industry work. These activities provide the necessary context for any envisioned reuse initiative, and also provide valuable models of processes for enabling reuse.

Examples of opportunistic reuse activities include the following.

- NASA ESE mission data systems regularly employ "clone and own" reuse. The practitioners rely heavily on experts with extensive knowledge of the system being ported/cloned. Examples of ESE mission data systems practicing this type of reuse are TSDIS, SeaWifs/OCTS, MLS, and QuikTOMS. In the near future, the developers of the GPM data system expect to reuse 60% of TSDIS system and thereby save 50% of the development costs. Some have suggested that the MODIS Rapid Response system is also a good model.
- The Software Engineering Institute at Carnegie-Mellon University, although primarily focused on strategic reuse, has defined a process for mining existing assets called "Options Analysis for Re-Engineering" which can be used to support opportunistic reuse.[22]
- NOAA undertook an effort to standardize coding conventions in part to make clone and own reuse easier.

Systematic software reuse incorporating product line concepts is employed at NASA, within the DoD, and in the commercial sector.

- Divisions at Goddard have independently undertaken reuse related initiatives. For example, the Flight Dynamics Division at Goddard realized 300% increase in reuse by applying OO techniques and a product line approach to a series of telemetry and

[22] J. Bergey et al., "Options Analysis for Reengineering (OAR): A Method for Mining Legacy Assets", CMU/SEI-2001-TN-013, http://www.sei.cmu.edu/publications/documents/01.reports/01tn013/01tn013figures.html#app-a.

dynamics simulators[23]. Also, the Flight Software Division has apparently incorporated reuse into its software development processes (based on the Software Productivity Consortium's product line engineering approach).

- The EOSDIS Core System arguably employed a product line approach to provide core components to the various DAACs. It appears that the expected cost savings were not realized by this approach.
- NASA sponsors some amount of research and cooperative agreements in software reuse. The Software Optimization and Reuse Technology Program at the West Virginia High Technology Foundation during 1995-2001 investigated domain engineering for optimization and reuse[24]. Two pilot projects focused on a product line approach for trend analysis and mission services respectively. The effort encountered a number of difficulties including a poor fit of product lines with types of systems being developed at Goddard and the NASA organizational/funding structures. A grant to Georgia Tech (Supporting Software Reuse, NAG 5-2226, 1995) investigated case-based retrieval of software artifacts as a means of capturing institutional memory and supporting reuse. The Repository-Based Software Engineering research program administered by JSC aimed to provide a repository and architectures that facilitate the selection, acquisition, integration, and reuse of software components. The repository is known as Electronic Library Services and Applications (ELSA). Closely related to this, the Reusable Objects Software Environment (ROSE) is being developed at NASA's Johnson Space Center to provide an economical and effective approach to reengineering and maintaining Flight Design and Dynamics systems.
- The SEI has focused significant attention on software product lines, and has published a book on this subject.[25] Included in the book are several interesting case studies, including an effort in which the National Reconnaissance Office contracted to Hughes/Raytheon to produce the core asset base, which would be reused and tailored to build multiple satellite command/control systems. This effort provides valuable lessons to NASA's ESE, because, in spite of proclaimed successes, it appears that organizational issues caused different programs to clone the core assets (which is counter to the product line approach). This effort also suffered from a lack of a documented architecture for the product line, again a valuable lesson for the ESE.
- The Software Productivity Consortium offers information and services related to reuse, including their Product Line Management and Engineering process, their Synthesis reuse methodology, and reuse capability assessments. NASA is an affiliate member and thus should have access to member-only information.

[23] "Impact of Ada and Object-Oriented Design in the Flight Dynamics Division at Goddard Space Flight Center". Software Engineering Laboratory Series, SEL-95-001, March 1995 [http://archive.adaic.com/docs/flyers/nasa.html]. See also F. McGarry, "Software Process Improvement in the NASA Software Engineering Laboratory", CMU/SEI-94-TR-22, [http://www.sei.cmu.edu/pub/documents/94.reports/pdf/tr22.94.pdf].

[24] See http://sort.wvhtf.org.

[25] P. Clements and L. Northrop, *Software Product Lines: Practices and Patterns.* Addison-Wesley, 2002.

- MacDonald Deitweiler Earth Observation Ground Systems Group is a component vendor that builds and sells “large” (up to 1 million LOC) components of remote sensing data systems. The company builds this software with reuse in mind, and employs such principles as low coupling, simple interfaces, and no reliance on COTS.
- Northrop Grumman has been applying product line methodologies and reference architecture concepts to build defense systems such as airborne radar systems and space sensors in less time with increased reliability.
- Commercial software product development companies such as Adobe Software rigorously employ product line concepts, maintaining strict control over a common set of core assets. Unfortunately, because of proprietary concerns of the companies involved, little specific information is available, but there are reports that show varied levels of success achieved by these efforts. For certain projects, Toshiba claims 60% reuse achieved and a 20-30% reduction in defects per LOC. Toshiba Software Factory reported 50% reuse over its product line in 1989 and increased productivity by 57%[26] by adopting an approach that promoted rewriting existing program modules in a generic form, replacing names of entities and relationships by more general terms. These generalizations called *presentations* were the basis for reuse and were propagated back through the levels of abstraction. Once a requirements level specification had been produced, designers tried to match it to an available presentation. If they succeeded, the designers traced through to the corresponding program level, converting generalized forms to a specific solution by making appropriate instantiations. Fujitsu has increased its on-time delivery rate for electronic switching systems software from 20% to 70%. And Hewlett Packard claims its time to market for individual projects was reduced by 43% and defects reduced 25%-75%. Since 1996 IBM has made asset-based development a core part of its enterprise-scale solutions, and claims numerous positive results including large improvements in delivery time.[27] However, software reuse success in the commercial world is not a given. A 1997 survey of 24 European companies showed 1/3 abandoned their reuse program because of poor results or an inability to make the program work.

Examples of open source reuse activities include the following:

- There are a number of open source software publishers that make the source code for entire applications available for reuse. The Open Channel Foundation (www.openchannelsoftware.org), offers more than 200 applications plus the NASA COSMIC collection. SourceForge (www.sourceforge.net) hosts over 36,000 projects; it offers tools for open source development (bug tracking, configuration management, document management, etc.) in addition to the code itself. FreshMeat (www.freshmeat.net) is a large catalog of Unix/Linux, Palm, and other cross platform, open source software. SlashDot provides news relevant to the open source

[26] M. Cusumano, "The Software Factory: A Historical Interpretation," *IEEE Software* (March 1989)pp. 23-30

[27] “Technical Reference Architectures”, P.T.L. Lloyd and G.M. Galambos; IBM Systems Journal, Vol. 38 No. 1, 1999.

development community. SourceForge, FreshMeat, and SlashDot are all owned by VA Software and its subsidiary, the Open Source Developer's Network.
- The National HPCC Software Exchange (http://www.nhse.org) created a "Repository in a Box" to facilitate the creation of reusable software repositories, and maintains a list of domain-specific repositories. The Computational Technologies Project (formerly the Earth and Space Sciences Project) at JPL and GSFC offers a catalog of analysis tools, parallel processing tools, and more (http://bryce.jpl.nasa.gov/catalog.pl?rh=3).

Additional initiatives not associated with any single approach include the following.

- NASA Goddard's Software Engineering Laboratory employs a bottom-up approach of capturing and reusing best practices (including policies, processes, tools, and training) sometimes referred to as a software experience factory. The SEL focuses on understanding, assessing, and packaging these practices (sel.gsfc.nasa.gov).
- There are hundreds of component and application repositories on the Internet that cover material of all types: public domain and proprietary, binary and source, code and documentation, freeware and shareware and traditional, etc. Examples in addition to those mentioned under the open source section above include the GNU project Web server, CNET's Download.com, the Comprehensive Perl Archive Network, Tucows, ZDNet Downloads, Netlib, the Numerical Algorithms Group's Numerical Library, and Pass the Shareware.
- The DoD has put significant effort into software reuse initiatives over the years. The DARPA-sponsored Software Technology for Adaptable, Reliable Systems (STARS) program, completed in 1996, included a heavy emphasis on reuse (http://www.asset.com/stars/) in general and product lines, architectures, and domain engineering in particular. In 1992 the STARS program successfully demonstrated the product line approach in three service projects involving the Army, Navy, and Air Force. The DoD Software Reuse Initiative, also completed in 1996, captured publications, lessons learned, successes, methodologies and more in the Reuse Information Clearinghouse (http://dii-sw.ncr.disa.mil/ReuseIC/) (may be accessible only from .mil domains). This effort was continued with the Central Archive for Reusable Defense Software (CARDS) program, jointly sponsored by the Air Force and NASA, which utilizes key STARS reuse technology in support of DoD and other government initiatives. Phase II of the CARDS program is a concerted DoD effort to transition advances in the techniques and technology of library-centered, domain-specific software reuse into mainstream DoD software procurements. The Library Operations Policies and Procedures (LOPP) serves as a comprehensive guide for operating and maintaining a reuse library. The Portable Reusable Integrated Software Modules (PRISM) Program demonstrated significant ROI using an architecture-based, product line approach to reuse.
- Academic research is investigating the use of formal methods and application generation to achieve reuse. For example, see http://liinwww.ira.uka.de/bibliography/SE/reuse.html.
- The National Association of State Chief Information Officers (NASCIO) is working with a group of states and vendors to iron out the details of a national repository of

software components. ComponentSource provides the infrastructure for the NASCIO component marketplace.

- Policy decisions can be used to drive reuse without specifying an approach. For example, the federal government in the 1990s said it would pay for only five states to develop a particular software application needed for child welfare systems; the other 45 states would have to reuse the code.
- NASA Goddard's Center for Software Engineering does not have specific initiatives/activities on software reuse, but does support process improvement and architecture activities into which reuse initiatives could be incorporated. For example, the Asset Management Group of Goddard's CMMI-based software process improvement initiative intends to maintain a database of GSFC software product characteristics in order to understand process metrics, encourage software reuse, and assist in identifying special expertise.
- The Workshop on Institutionalizing Software Reuse offers annual workshops and a variety of position papers related to this topic (http://www.umcs.maine.edu/~ftp/wisr/wisr9/final-papers/TOC.html).
- The IEEE in 1994 signed with the Reuse Library Interoperability Group (RIG) to cooperatively develop standards for setting up and linking libraries of reusable software.

4.2 Reference Architecture

There are numerous activities related to ESE reference architectures. Much of this work is being performed within standards organizations, which typically use reference architectures to partition and coordinate the work of subcommittees.

- The OGC is working mostly on detailed GIS standards, but its OpenGIS Service Architecture identifies services (Access Services, Display Services, Imagery Manipulation Services, etc.) that serve as a reference architecture for organizing standards development efforts. It draws from the Open Services Environment model (ISO 19101) and Reference Model for Open Distributed Processing. As an aside, twenty vendors offer more than seventy products claimed to be OpenGIS implementations.
- The Federal Geographic Data Committee (FGDC) is the focal point for a number of related activities. The Data Framework defines (among other things) geospatial data themes used to facilitate collection, use, and maintenance of commonly needed geospatial data for the National Spatial Data Infrastructure. It also defines procedures, technology, and guidelines that provide for integration, sharing, and use of data within those themes. The FGDC Standards Reference Model defines categories of geospatial standards; the process standards subcategories therein relate to a top-level functional architecture. The Geospatial Applications & Interoperability working group builds on the Digital Earth Reference Model (now called the Geospatial Interoperability Reference Model) to promote standards for seamless access to distributed data.
- The Global Spatial Data Infrastructure is promoting complementary policies, common standards and effective mechanisms for the development and availability of

interoperable digital geographic data and technologies to support decision making at all scales for multiple purposes. While it does not appear to have any reference architecture initiatives per se (it draws from the FGDC for this need), the *SDI Implementation Guide: a Cookbook to support the GSDI* is a practical guide that fulfills some of the objectives of a reference architecture.

- The Global Grid initiative is defining the infrastructure needed to provide scientific and engineering computing communities with access to large-scale, distributed computing and storage resources. Participants include NASA (through the Information Power Grid program), DOE (ANL), NSF, IBM, Microsoft, Cisco, USC, the NCSA Alliance, the National Partnership for Advance Computing Infrastructure, and others. The Open Grid Services Architecture is an evolving reference architecture that unifies Web and Grid services. The Globus Toolkit provides a set of utilities to facilitate connecting to the Grid.
- The American National Standards Institute, National Committee for Information Technology Standards, Geographic Information Systems Technical Committee (L1) does not appear to have significant reference architecture initiatives except in connection with its role as the Technical Advisory Group to ISO/TC211 (see below).
- The ISO Technical Committee on Geographic Information / Geomatics (TC 211) is developing geospatial standards in five areas, including three related to reference architectures: “Framework and Reference Model” (ISO/DIS 19101), “Profiles and Functional Standards” (ISO/TR 19106 and 19120), and “Geographic Information Services (ISO 19119). The reference model describes the environment within which the standardization of geographic information takes place, the fundamental principles that will apply, and the architectural framework for standardization. The reference model defines and relates all concepts and components needed for this standardization. Structured within information technology standards, the reference model will be independent of any application, methodology, and technology. ISO 19119 (reused by OGC) includes a taxonomy of geospatial services that also serves as a more detailed reference model.
- The ISO has other related activities outside of the geospatial domain, such as the Reference Model for Open Distributed Processing (ISO/IEC IS 10746), which defines viewpoints, language, functions, and transparency prescriptions needed to specify open distributed processing systems, and the venerable Open Systems Interconnection reference model. The European Commission has sponsored a GIS interoperability project based on the RM-ODP titled “Distributed Geographical Information Systems - Models Methods Tools and Frameworks” (DISGIS, Project ESPRIT 4).
- The IEEE has related activities outside of the geospatial domain, such as the Metadata Reference Model and the Reference Model for Opens Storage System Interconnection.
- The Open Group has related activities outside of the geospatial domain, such as the Data Management Reference Model (G505), the Distributed Transaction Processing Reference Model (G120), and the Systems Management Reference Model (C207). The Open Group also promotes the Architecture Description Markup Language (ADML) for the capture and interchange of architecture descriptions.

- The CCSDS sponsors the Open Archive Information System (CCSDS 650.0) recommendation, which was created in response to ISO TC20/SC 13 to provide a reference model to facilitate discussion and comparison of archives.
- The OMG has defined an Object Management Architecture that serves as reference architecture for distributed computing services as part of the Common Object Request Broker Architecture. Somewhat related, the OMG's Model Driven Architecture is a new way of writing specifications and developing applications that segregates platform-independent and platform-dependent models. Also, OMG has responsibility for the Unified Modeling Language, which is a graphical language increasingly used to define architectures.
- The Interoperability Clearinghouse has a number of related activities including a domain architecture working group, an architecture assurance methodology effort, and an architecture "collaboratory".

Various government organizations are also working on reference architectures, typically to improve interoperability of systems developed by different contractors:

- NASA GSFC MO&DSD defined the Renaissance Ground Data Systems Architecture to provide a reference architecture for mission ground data systems (Renaissance is an acronym for REusable Network Architecture for Interoperable Space Science, Analysis, Navigation and Control Environments). This work has subsequently been moved to the OMG Space Domain Task Force, Space Object Technology Group. The Common Modeling Infrastructure Working Group aims to organize a framework and determine standards to improve climate model interoperability. Similarly, the Computational Technologies Project has initiated (and funded through a CAN) the Earth System Modeling Framework, which aims to define a specific software architecture to improve model component interoperability and reuse. ESTO's AIST program appears to have several related thrusts (e.g., Earth Science Interoperable Data & Services Framework, and Data Product Planning & Scheduling). NASA also participates in many of the reference architecture activities in standards organizations mentioned above.
- The National Imagery and Mapping Agency (NIMA) has documents that serve the same purpose as a reference architecture. For example, the NIMA Implementation Plan for the DoD Joint Technical Architecture establishes processes to ensure that JTA interoperability standards are incorporated into all acquisition and development processes. Also, the high level USIGS architecture documents effectively establish a reference architecture. See, for example, the USIGS Objective System Architecture Description (which is the USIGS vision for the 2005-2010 timeframe), the USIGS Common Object Specification (which defines interfaces, data types, and error conditions to prevent redundant or non-interoperable specifications), and the USIGS Interoperability Profile (which identifies key interfaces and related data interchange standards that define the minimum requirements for access and connectivity among applications). NIMA also hosted the multi-service Global Geospatial Information and Services (GGIS) Integrated Product Team, which defined a top-level architecture aimed at improving global production and dissemination of geospatial information (particularly maps for warfighters).

- DARPA's Domain Specific Software Architecture program aims to demonstrate the value of architectural concepts to allow early validation and iterative development, decrease risks, and decrease costs. Sample results for real-time systems include a 20:1 reduction in requirements definition time; the process was based on the STARS domain analysis process (Diaz) and Feature-Oriented Domain Analysis (Cohen/SEI). DARPA's Advanced ISR Management research program aims to define a reference architecture and technologies for improved demand-based data acquisition and processing. DARPA's Intelligent Integration of Information (I^3) program aims to significantly reduce the time and complexity involved in sharing and integrating heterogeneous information on large-scale, widely distributed networks; one facet is the development of an I^3 reference architecture to serve as a vision for the program.
- A variety of DoD organizations including the Air Force and Army (with support from the SEI) have developed domain specific software architectures (for flight simulation and movement control, respectively) with the goal of making it easier to generate specific instances of these systems. The SEI cites the Tektronix oscilloscope architecture (based on pipes and filters) as an early example of a domain specific software architecture (DSSA) for a product line, and the Sematech Computer-Integrated Manufacturing Framework as a more recent example of a DSSA. A DSSA is essentially a reference architecture with a software focus.
- DISA has related activities as part of it Common Operating Environment initiative, which defines a foundation for building open systems. For example, the COE includes profiles of geospatial data standards and reference datasets. The Defense Information Infrastructure and the Global Information Grid initiatives also aim to define a set of interoperable capabilities, but on a more globally distributed basis, and will interface standards (vs. compatible implementations) as the basis for interoperability.
- The Air Force Portable, Reusable, Integrated Software Modules (PRISM) program extracted a command center reference architecture from exemplar systems. The goal of PRISM was to reduce acquisition costs through software reuse and standardization of architectures. An ROI of 377% ($7M investment resulted in $26M savings across five PRISM programs) is claimed[28].
- Other government agencies have embarked on enterprise architecture efforts that provide specific guidance within their domain. For example, the VHA Enterprise Architecture initiative defines a standards profile and key components to guide and facilitate acquisition and development of interoperable systems. The USPTO claims a cost savings of $32M/yr (15%) from their enterprise architecture effort.[29] HUD has a concerted enterprise architecture effort, and built a Web-based tool (the Enterprise Architecture Management System) to manage the architecture. The Department of Agriculture is taking a more bottoms-up approach to their enterprise architecture, using processes that may be applicable to a NASA reference architecture effort.
- The DOE Common Component Architecture Project aims to define a generic component architecture that supports abstractions necessary for high-performance computing.

[28] http://sunset.usc.edu/GSAW/GSAW99/pdf-presentations/breakout-2/boehm.pdf

[29] http://www-trm.itsi.disa.mil/uspto_case_study_041701.pdf

- The NASA Data and Information Access Link is a Web-based HDF file server (previously known as “DAAC-in-a-box) that goes beyond a reference architecture into a reusable reference implementation.
- The Ground Systems Architecture Workshop explores issues and potential for consensus in software architectures for spacecraft ground systems. While NewDISS is concerned with a different segment of the overall system, the problem set is similar technically.

There is also some related activity in industry and the computer science community:

- IBM established the Enterprise Solutions Structure initiative to facilitate the construction of systems from architecturally-compliant assets and stop the growth of “silo” solutions. IBM will use reference architectures as the basis for its asset-based approach to solution development. System architects Lloyd and Galambos at IBM assert that defining a constrained set of reference architectures is not only feasible, but mandatory for large-scale development.[30]
- There is active work combining the concepts of patterns with software architecture. See, for example, Pattern-Oriented Software Architecture, Volume 1: A System of Patterns (Buschmann et al.).

Other related activities without specific reference architecture content include the following:

- There are a number of efforts that define architecture frameworks that provide useful guidance for creating reference architectures, generally by defining a set of perspectives to describe the architecture. Examples include CMU’s Implicit Structuralism, the C4ISR Architecture Framework (operational, technical, system), Kruchten/Rational Framework (design, process, implementation, deployment, use case), AF Integrated C2 System (capability, data, distribution, security, construction), the Zachman Framework (36 views), and The Open Group Architectural Framework (foundation, common, industry, organization). IEEE 1471 is an all-encompassing recommended practice for architecture specification. OMB Circular A-130 includes enterprise architecture requirements for Federal agencies. The CIO Council and NIST have lead roles in defining the Federal Enterprise Architecture Framework, which provides important guidance for the development of enterprise architectures (per Circular A-130) and, by extension, reference architectures.
- The ESIP Federation has an Interoperability Standing Committee that is defining the “System Wide Interoperability Layer” or “Federation Interactive Network for Discovery” for distributed catalog and data access, but does not appear to have any reference architecture initiatives.
- The Open-Source Project for a Network Data Access Protocol (OPeNDAP), an outgrowth of the Distributed Oceanographic Data System (DODS) project, is a non-profit corporation established to develop and promote software that facilitates access

[30] “Technical Reference Architectures”, P.T.L. Lloyd and G.M. Galambos; IBM Systems Journal, Vol. 38 No. 1, 1999.

to data via the network. By implementing the data access protocol and adopting the implied network data access architecture, data-level access across roughly a dozen systems has been demonstrated.

- The Simulation Interoperability Standards Organization focuses on facilitating interoperability and component reuse, but does not have any reference architecture activities per se.
- The Joint Steering Group on Spatial Standardization and Related Interoperability has performed a survey of potential areas of coordination among its members (primarily standards organizations), but has not yet tied this to a reference architecture.
- The Digital Geographic Information Working Group (sponsor of the DIGEST data format standards) does not appear to have any reference architecture activities.
- The International Steering Committee for Global Mapping does not appear to have any initiatives related to reference architectures.
- The Digital Libraries Initiative (incl. Project Alexandria) is a multi-agency research program to create large knowledge bases, the technology needed to access them, and the means for improving their usability in a wide range of contexts.

5 Evaluation of Alternative Approaches

5.1 Evaluation Criteria

To provide a basis for evaluating reuse methodologies and reference architecture alternatives for potential applicability to the SEEDS environment, a series of primary evaluation criteria and sub-criteria were established. These primary criteria reflect the core goals of the SEEDS initiative. Figure 4.1-1 shows the primary and sub-criteria used in the evaluations. Each criterion was then evaluated subjectively based on the System Engineering Team's own experience and on opinions from the community. The results of the evaluations are described in section 4.4.

Primary Criteria	**Rationale for Criteria**	**Sub-Criteria**	
Reduction of cost of supporting future missions, science, and applications	Flat budgets predicted for ESE information systems; current missions consume most of that projected budget	**Reuse** Time to Market Development Efficiency System Maintenance	**Reference Architecture** Proposal Prep and Evaluation Competition Development Efficiency Integration Effort
Increased flexibility and responsiveness to new missions, science, and applications	Current large development efforts require rigid requirements control and can't adapt quickly to changing science and technology	New Requirements New Science Support Technology Exploitation	
Increased effective and accountable community participation	Community participation increases domain expertise and should increase productivity	None	
Suitability for flight mission needs	Approach must be able to address the schedule-driven emphasis of flight missions	Cultural Needs Organizational Needs	
Suitability for ESIP-type needs	Approach must be able address the ESIP's emphasis on	Cultural Needs Organizational Needs	

	innovation	
Relative investment cost	Tight budgets require cost-effective implementation and maintenance/support of an approach	Process Support/ Coordination Generalization and Documentation

Figure 5.1-1 Criteria used to evaluate software reuse and reference architecture approaches

5.2 Reuse Alternatives

With input from the NASA community and extensive research of current industry as well as government reuse practices, the study team identified the following reuse alternatives:

- Status Quo
- Single System Development with Reuse ("Improved Clone & Own")
- Open Source Software Development
- Encapsulated Services
- Product Lines

An overview of each of the alternatives is presented in this section. The results of evaluating each of the alternatives are summarized in Section 0 "2.2 Community Opinion" with additional details provided in Section 0 "5.4 Approach Survey and Analysis".

5.2.1 Status Quo

The default alternative for NASA is to continue employing the current mix of *ad hoc* "clone & own" practices and use of a single centralized contractor.

Ad hoc "clone & own" refers to the practice of developing a new mission system by copying an existing fully functional system and modifying it to fit the needs of a new mission ("cloning"). The new "cloned" system is totally independent of the original one and its maintenance and support are the sole responsibility of the new mission ("owning").

The use of a single centralized contractor refers to the practice of hiring a contractor to develop and maintain all or parts of a new mission system.

5.2.2 Single System Development with Reuse ("Improved Clone & Own")

An alternative reuse approach is for NASA to extend the current *ad hoc* "clone & own" practice with supporting processes and tools to allow it to be used more easily and with better success by more groups. Potential supporting processes can target mechanisms for

- Enabling developer groups to identify, locate and understand available systems and/or artifacts that can be cloned
- Upgrading or creating new documentation on the available systems
- Providing developer groups with easy access to system experts for extended consultation and guidance

The “improved clone & own” approach still requires each cloned system to be maintained as a separate system. In order to minimize the risk of continuously cloning old and/or technically obsolete systems, this approach needs to address technology upgrades and infusion issues.

5.2.3 Open Source Software Development

With open source software development, selected system components and/or systems are collaboratively developed and updated by developers across projects. Developers are free to check components out of a code repository and adapt them to their own systems. Developers contribute back to the repository by submitting updates, bug fixes and enhancements patches. A repository-control authority determines which developer contributions get incorporated into the code base.

The independent peer review and problem debugging resulting from the collaborative environment of open source development may lead to better product quality. However, reliability of available components often mission-critical applications cannot be guaranteed given that these components are built by a variety of developers for a variety of purposes. Furthermore, long-term maintenance of open source code can be problematic because developers may lose interest in maintaining their components, or may be hesitant to support change requests form other groups.

5.2.4 Encapsulated Services

The Encapsulated Services alternative involves wrapping existing systems or components with network-accessible wrappers (e.g. HTTP accessible, Microsoft .Net service, Java RMI). Services can be invoked by others at NASA for access and use as needed within their systems.

This approach may require a policy shift at NASA as the groups that own the wrapped components become service providers. Maintaining services and ensuring they meet pre-defined quality of service levels may cause the service providers to experience net resource drains unless service “customers” help pay for the operations. Such an approach will also require a significant technology development which may be too costly and time-consuming for NASA to implement in the short-term.

5.2.5 Product Lines

The Product Lines alternative is a well-established systematic reuse approach which centers around the process of identifying, creating, maintaining and evolving common core assets which can be easily integrated to build sets ("lines") of related new systems ("products"). The Product Lines approach makes a clear distinction between core asset development and product development.

- Core asset developers build the core assets either from scratch or by mining existing systems and tailoring them for reuse. Variation points are built into the architecture and the core assets to allow for ease of tailoring of product lines. The maintenance of the core assets is the responsibility of the core asset developers. The high initial investment needed to build the core asset base requires reuse to be done multiple times to realize cost savings.
- Product developers must understand how to use the core assets, and in particular the variation points built into the architecture and the core assets to tailor their products. Product developers cannot modify the core assets.

With this distinction between core asset development and product development, the application of Product Lines at NASA would require a new organizational structure that is not mission-driven.

5.3 Reference Architecture Alternatives

There are a variety of approaches that NASA could use for an ESE reference architecture effort. These approaches differ from each other in three primary dimensions:

- The level of specificity in the architecture. For example, the architecture could simply provide textual descriptions of interfaces, or it could invoke formal interface standards.
- The level of granularity in the architecture. For example, the architecture could define only the top dozen or so major components and interfaces, or could break each of those down into (potentially procurable) smaller components or internal interfaces.
- The focus of the architecture. For example, the architecture could be limited to communications and infrastructure services to achieve data-exchange capability, or it could include domain-specific services to enable plug-and-play applications.

The choice of approach depends primarily on how the reference architecture is intended to be used. For example, the architecture could simply be offered as helpful guidance, or could be incorporated into contracts for mandatory compliance. Also, it could be intended to promote compatibility at the subsystem level or software module level, thereby enabling reuse at either the subsystem level or software module level.

From a process and funding perspective, there are also a variety of options. For example, NASA could withdraw its current support for activities related to reference architectures, it could leave its current support unchanged, it could re-prioritize and shift funding, or it could define and fund a new initiative. These options will not be considered in depth until after the merits of a reference architecture itself have been considered.

To simplify the analysis, the study focused on the first two dimensions. The following definitions are used for levels of specificity:

- Notional. In this approach, NASA would support development of a reference architecture that would provide only a high-level decomposition of the functional components of an ESE data system. The architecture would define a common structure and terminology that could be used, for example, to facilitate communications or clarify organizational responsibilities, but would not contain enough detail to ensure interoperability of separately developed software. It would attempt to unify individual activities already underway into a focused initiative.
- Concrete. In this approach, NASA would support development of a reference architecture that would define functional components, abstract service invocation models (e.g., call vs. message), ancillary services (e.g., directory services), and a limited set of standards. The architecture would provide enough detail to facilitate interoperability of separately developed subsystems (perhaps through gateways or other translation mechanisms), but would not itself ensure interoperable implementations. The granularity would be at least at the second level (i.e., one below the top dozen or so major functional components). Participation in standards activities would be increased to help close identified gaps in the reference architecture.
- Specific. In this approach, NASA would add detail beyond a concrete reference architecture to include specific service invocation mechanisms and a comprehensive technical architecture (i.e., a detailed list of services, applicable standards, and recommended products). The architecture would provide a level of specificity that enables interoperation of separately developed software subsystems, components, or modules. It would be used not only to drive standardization efforts, but also to select from (and qualify) the results of those efforts.

The following definitions are used for levels of granularity:

- Coarse. In this approach, functions would be decomposed only to the level of a subsystem (such as an ingest subsystem, processing subsystem, or access/dissemination subsystem) with the goal of enabling reuse of entire subsystems.
- Medium. In this approach, functions would be decomposed to the level of a component such as a data catalog, or job dispatcher with the goal of enabling reuse of packaged applications.
- Fine. In this approach, functions would be decomposed to the level of an individual software module, with a goal of enabling reuse of code modules.

5.4 Approach Survey and Analysis

The study team had a variety of stakeholders from the ESE community evaluate the alternative approaches to reuse and reference architectures. The results of the evaluation are shown in “2.2   Community Opinion” above. In addition, the study team consolidated the written comments from the evaluation, which discussed the structure of the evaluation (e.g., the evaluation criteria and alternatives), qualified the evaluation results, and highlighted a variety of concerns and recommendations. The consolidated comments are provided in the following sections.

5.4.1 Reuse

Comments about Reuse Evaluation Criteria

Feedback about the evaluation criteria was positive as it was felt that the criteria were well-aligned with the overall objectives of SEEDS. There were suggestions for considering the following as additional evaluation criteria:

- Technical risk, as it is often the main driver in specific instances.
- The cost impact on key resources such as creative personnel, decision makers and managers.

Comments on Reuse Options

Various individuals from the community expressed their concerns about continuing with the status quo as it is costly and stifles innovation.

Product lines ranked low in the evaluation as it was felt that they exhibit a high level of technical risk since assets produced for the purpose of reuse tend to be new and often not robust. The negative experiences of some participants with the development of the ECS system using the product line approach also contributed to the overall negative assessment of this option.

As for clone & own, members of the community cautioned that the improved approach should be a more efficient rather than a more controlled version of the current *ad hoc* practice. It was also stressed that the evolution of core competencies is integral to the success of the approach.

In terms of the community’s response to the open source approach, there was some concern that there may not be enough of a critical mass for this approach to achieve its potential within NASA, and that its non-deterministic nature may make it impossible to properly estimate and allocate cost and schedule.

A similar concern was expressed in the case of the encapsulated services approach. Skepticism regarding people and organizations supporting such an approach was expressed.

A combination of the presented alternatives was proposed, consisting of institutionalizing the clone & own practice while borrowing and incorporating open source components as well as using services provided by some organizations as encapsulated services.

Recommendations

The community was instrumental in supplying the study team with implementation-specific recommendations, including:

- Focus on reusable tools (with the rationale that scientists care more about the use of tools that work rather than the challenge of developing data systems)
- Focus on general software which has a broader opportunity for reuse as opposed to discipline and standard-specific software which has limited reuse opportunities
- Identify and focus on systems and components that have a high potential for reusability.

The community also stressed the need for metrics, well-defined interfaces and incentives to compensate systems or components that are repeatedly used (cloned/called).

Concerns

As expected, some members of the community expressed their concerns as to whether reuse will inhibit competition and hurt evolution at NASA. Others felt that the reuse of science software is quite variable and that reuse in general was not a solved problem in Computer Science. There was also the concern that this study is focusing more on the supporting systems rather than the science systems.

5.4.2 Reference Architectures

General Comments

The stakeholders commented that reference architectures are useful for the following reasons:

- They allow assessing the data system approach against SEEDS goals;
- They permit assessment of component utility for the user;
- A coarse architecture can encourage low coupling and increase flexibility; and
- Portions of a reference architecture (specifically the data model, interface definitions, and self-describing data formats) were seen as useful parts of ECS even by ECS critics.

Several stakeholders also noted that what is needed is not *a* reference architecture, but *a set of* reference architectures. The architectures may vary in detail or applicability to different environments, or could be used to characterize existing systems (to support the clone & own approach to reuse).

Finally, several stakeholders noted that reference *implementations* can be useful in addition to a reference architecture, primarily for the purpose of demonstrating compatibility or conformance with the architecture, but also to speed adoption of the architecture.

Concerns

Stakeholders expressed concern that a reference architecture might reduce innovation, community participation, and technology infusion. By definition, an architecture introduces constraints…and more specific architectures have more constraints. This concern is the main driver behind relatively poor rating of the "specific" architecture alternative. A number of stakeholders stated directly that specific reference architectures are not appropriate for the mission-success community.

Stakeholders were also concerned that the relationship between developing a reference architecture and advancing the ESE science goals were not clear. One stakeholder noted that, in particular, easy access to data at the proper granularity is what scientists really care about. The study team subsequently clarified the relationship by directly linking reference architectures to the ESE strategic plan (see Section 0 "1.1 Motivation" above) and focusing on using reference architectures to facilitate software reuse.

5.5 Survey Form

The following form was used to solicit community input regarding alternative reuse and reference architecture approaches.

Evaluator Information

Name:

Organization:

Current Activity:

Related Experience:

Job Focus: Choose oneÉ

Email:

Software Reuse

Option Criteria	Status Quo Reuse	Improved Clone & Own	Open Source	Service Encapsulation	Product Lines
1. System cost savings					
2. Flexibility & responsiveness					
3. Increased effective & accountable community participation					
4a. Suitability for ESE Mission Environment					
4b. Suitability for ESE Science/Applications Environment					
5. Investment cost					

Reference Architecture (Specificity)

Option Criteria	Status Quo Architecture	Notional	Concrete	Specific
1. System cost savings				
2. Flexibility & responsiveness				
3. Increased effective & accountable community participation				
4a. Suitability for ESE Mission Environment				
4b. Suitability for ESE Science/Applications Environment				
5. Investment cost				

Reference Architecture (Granularity)

Option Criteria	Coarse	Medium	Fine
1. System cost savings			
2. Flexibility & responsiveness			
3. Increased effective & accountable community participation			
4a. Suitability for ESE Mission Environment			
4b. Suitability for ESE Science/Applications Environment			
5. Investment cost			

6 Cost Savings Sensitivity Analysis

In order to gain a better understanding of potential reuse cost savings, the study team performed a cost savings sensitivity analysis using the Poulin/Caruso Reuse Metrics Model, developed at IBM in 1992[31]. It is important to note that in order to run this model, the study team used basic estimates for the model parameters (as opposed to parameter

[31] See *Measuring Software Reuse: Principles, Practices, and Economic Models* by Jeffrey Poulin, 1997.

values directly extracted from various ESE systems). As such, the analysis performed by the team did not reflect actual ESE experience. However, it still provided the team with some useful insights on potential cost savings that can be achieved as a result of a reuse initiative.

The section below provides an overview of the model followed by the results of the team's cost savings sensitivity analysis.

6.1 Overview of Model

The Poulin/Caruso model essentially calculates the return on investment of an organization as a result of a systematic reuse effort. In the case of NASA, the cost model requires the following inputs for each of the eight[32] known future missions:

- It is assumed that 20% of a system's code is unique to that mission[33].
- The average development cost of a single LOC ($/LOC), for which we assume of $100/LOC based on industry averages.
- The average number of errors per KLOC (errors/KLOC) and the estimated cost of each error ($/error): These values are used to estimate system service costs. The study team used values recommended by the Poulin/Caruso model as 0.5 errors/LOC and $10,000/error.

    - The Relative Cost of Reuse (RCR), for which we assume a value of 0.2 as recommended by Poulin. RCR represents "the ratio of the portion of the effort that it takes to reuse software to the cost normally incurred to develop it for one-time use". For example, if a piece of software is reused for only 20% of the cost of new development, RCR=0.2. The effort of reusing software includes the efforts to locate, evaluate and integrate a piece of code into an application.
    - The Relative Cost of Writing for Reuse (RCWR), for which we assume a value of 1.5 as recommended by Poulin. RCWR represents "the ratio of the portion of the effort that it takes to develop reusable software to the cost of writing it for one-time use". For example, if it costs an additional 50% effort to develop reusable code then RCWR=1.5. The effort of writing for reuse includes the efforts of domain analysis, creating a more generic design, providing additional documentation, etc.
    - The estimated/desired percentage of reusable code provided by each mission system for future mission systems, for which we assume a value of 5% based on perceived budget constraints.
    - The estimated/desired percentage of code reused by the mission data system, for which we assume values ranging from 30% to 60%.

Based on these inputs, the following outputs are calculated for each mission data system and consequently cumulatively for NASA:

[32] List provided by the short-term standards SEEDS study team.

[33] In the form of mission-specific requirements, algorithms, etc.

- Reuse Cost Avoidance (RCA): RCA represents the benefit of reusing software by estimating the money each mission and cumulatively NASA did not have to spend to develop new software. RCA is calculated as the sum of the Development Cost Avoidance (cost avoided during the development phase) and the Service Cost Avoidance (cost avoided during the service phase)[34]. RCA varies with the reuse percentage and the Relative Cost of Reuse.
- Additional Development Cost (ADC): additional costs incurred by developing reusable software. ADC varies with the percentage of code written for reuse and the Relative Cost of Writing for Reuse.

As an additional input, the study team estimated that a certain percentage of code can be mined from the systems currently available across NASA. The cost of mining these assets is also included in the analysis based on an RCWR value of 1.5 (i.e. it costs 50% more to mine existing assets in order to make them reusable by other mission data systems).

The percentage savings expected to be achieved by NASA as a result of a reuse effort is calculated by subtracting the initial mining cost and the Additional Development Cost from the Reuse Cost Avoidance and dividing that by the cost of developing eight mission data systems without reuse.

6.2 Cost Savings Sensitivity Analysis Results

The first step in the cost analysis consists of calculating any reuse savings expected to be reaped by continuing with the current *ad-hoc* clone & own practice. The cost savings model is next applied to the improved clone & own option in order to estimate the additional savings that will result from enhancing and institutionalizing the currently *ad-hoc* practice. The additional savings are then compared to the ones expected to be achieved using the more formal approach of product lines.

6.2.1 Current Ad-hoc Clone & Own

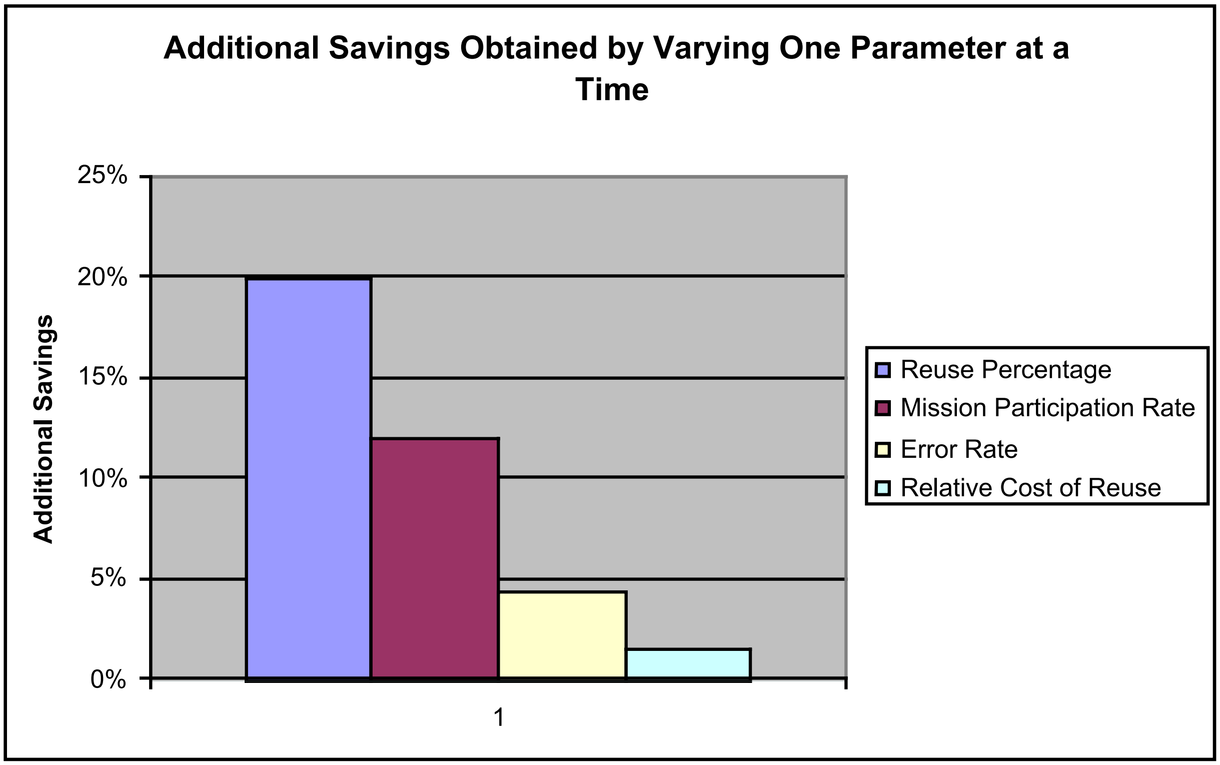
Applying the model to the current *ad-hoc* clone & own practice, total savings for NASA after 8 missions are estimated at 12%. This estimate is based on the assumptions that

- Every other mission will build its system using the clone & own approach (i.e. mission participation rate is 50%).
- The reuse level of the participating missions remains at a constant 30%.
- The participating missions will not invest in writing any of their software for reuse by others (hence, explaining the constant reuse level across all reusing missions).

The cost model shows that:

[34] Service Cost Avoidance is needed because reusable software tends to have up to ten times better quality than software developed for one-time reuse, which implies that a reusing mission is expected to eliminate a significant maintenance cost.

- Additional savings of 20% can be achieved by increasing the reuse level to 80% (without modifying any of the other factors)
- Additional savings of 12% can be achieved by increasing the mission participation rate to 100% (without modifying any of the other factors)
- Additional savings of 4% can be achieved by lowering the error rate to half its assumed value (without modifying any of the other factors)
- Additional savings of about 2% can be achieved by lowering the Relative Cost of Reuse to 0.1 (without modifying any of the other factors)



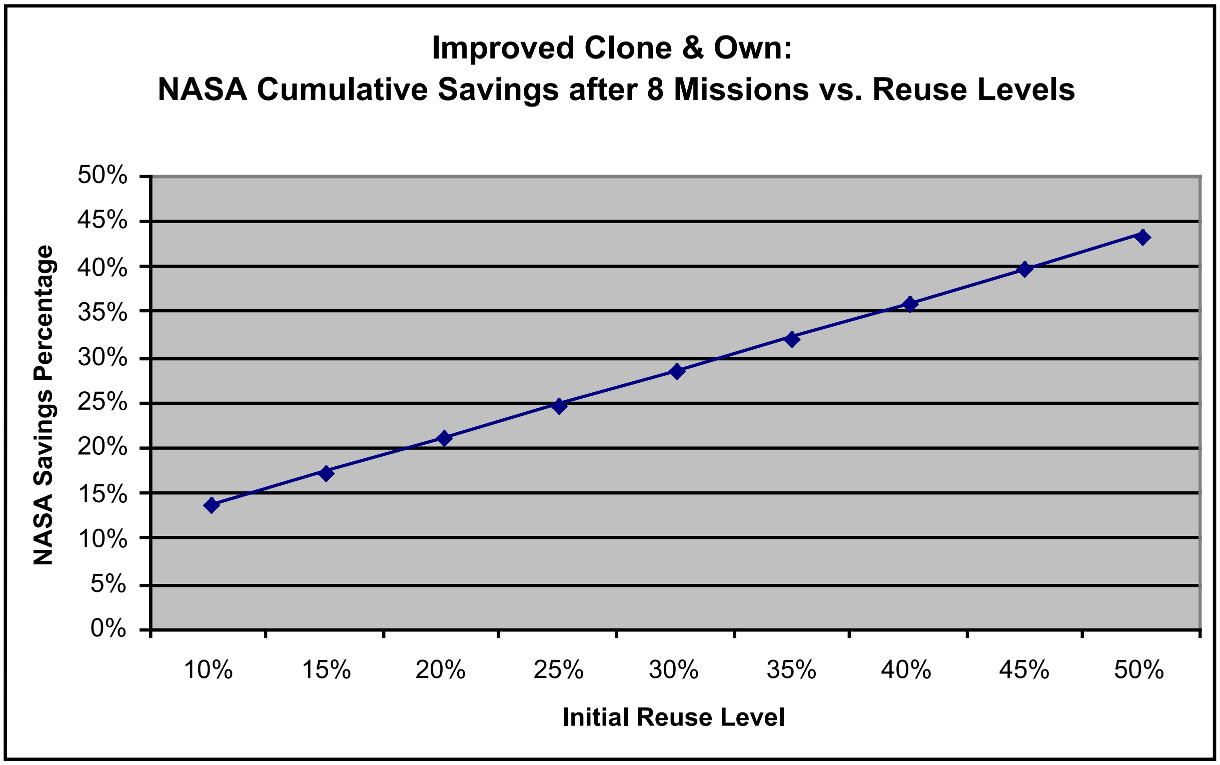
The figure above suggests that an institutionalized reuse process should start by focusing on improving the reuse percentage of participating missions while increasing the mission participation rate. The improved clone & own reuse option attempts to achieve that as described next.

6.2.2 Improved Clone & Own

With improved clone & own, it is expected that the reuse process will include a preparation step where existing systems are mined to produce reusable code for future systems. This mining step is then likely to increase the initial reuse level from 30 to 40%. As with any institutionalized reuse process, it is implied that all future missions will participate in the process by reusing available code and by contributing back to a reusable code repository. With the assumption that each mission can contribute 5% of its code back to the repository, the reuse level is expected to increase with every mission (except in the case of concurrent missions which are unlikely to be able to leverage each other's contributions).

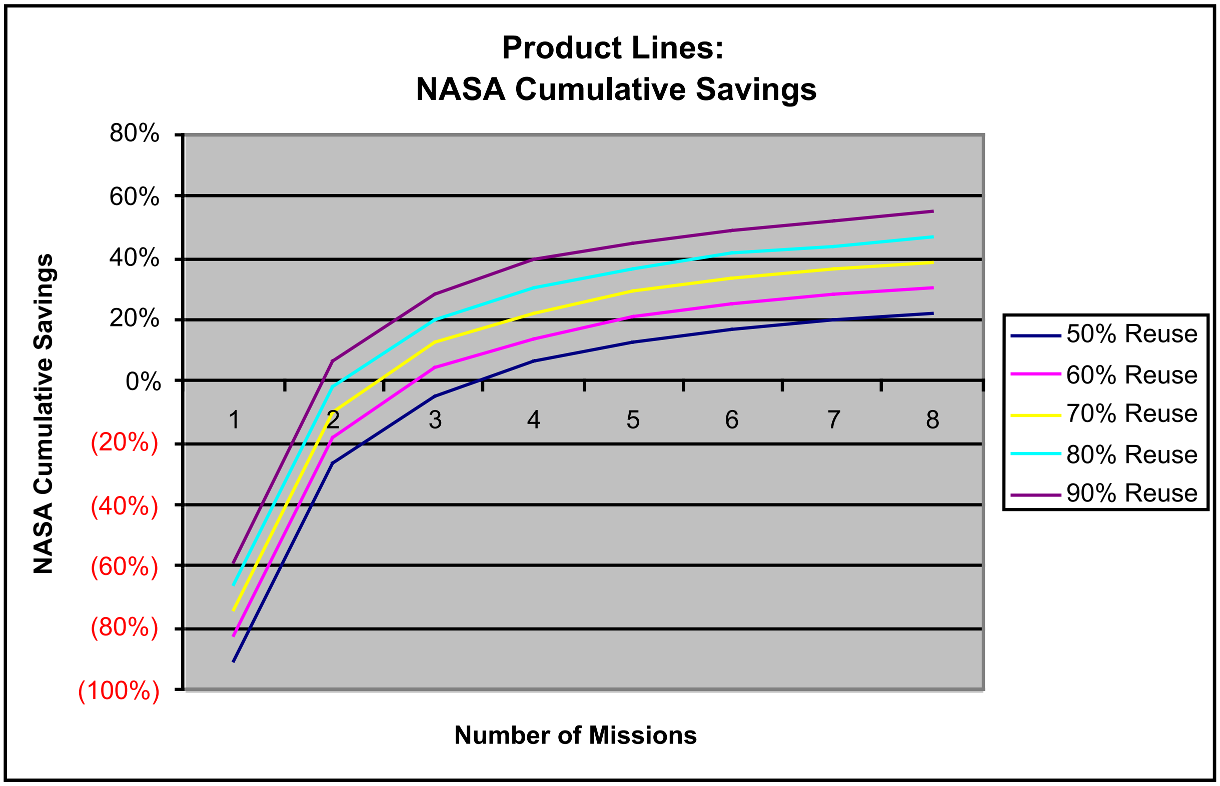
Running the cost model with RCWR = 1.5 and RCR = 0.2 (both values recommended by experts in the field), the analysis shows that improved clone & own option can provide NASA with total savings of 36%, translating into an additional 27% savings over the current *ad-hoc* practice. The money saved can be used for additional science support for participating missions.

The figure below shows that considerable savings can still be achieved even if the initial reuse level is lower than the assumed 40%.



6.2.3 Product Lines

A similar analysis is performed on the product lines option. Higher reuse levels are expected in this case since the product lines approach is based on the fact that a core asset base specifically designed and developed for reuse is used by all missions. However the development of this core asset base requires a considerable up-front investment, which in turn implies that significant savings can only be achieved when reuse is done successfully by all missions. The figure below shows that, after eight missions, the product lines approach can provide overall savings as high as 55% for a reuse level of 90%. However, with lower reuse levels and fewer missions, the savings percentage is considerably lower.



7 Reuse Process and Next Steps

7.1 Reuse Process Characterization

A good reuse process is needed to avoid the common causes of failure in reuse initiatives. In a 1997 survey of two dozen European companies, roughly one-third abandoned the reuse program because of poor results or an inability to make it work[35]. Similarly, a survey by the Standish Group shows that 70% of companies surveyed reported failures with their first reuse application development effort, although success rates improved substantially in subsequent efforts.[36]

The study team conducted breakout sessions at the SEEDS public workshops to facilitate community definition of processes to enable software reuse. In addition, the study team reviewed reuse literature and captured guidance that was relevant to the ESE environment.

[35] Morisio, M. & Tully, C. & Ezran, M. Diversity in Reuse Processes. IEEE Software. July/August 2000.

[36] “Hidden costs of software reuse”, InformationWeek, 1998

The first step was to identify guiding principles, contributing factors, program strategies, technical strategies, and evolutionary approaches. These are captured in the following sections with minimal vetting or prioritization to retain the essence of the original community guidance.[37] From this input, the study team synthesized a straw process, which includes a definition of reuse activities, information and control flows, supporting mechanisms, and organizational roles and responsibilities. This straw process is intended to be used as a starting point for further definition and refinement by the community.

Most of the input in the following sections was gathered at the Second SEEDS Public Workshop. Recall that community input from the first workshop indicated that two different reuse approaches would be appropriate: improved clone and own for mission-critical environments, and open source software for mission-success environments. With this in mind, the community input below has been kept separate for these two environments. Each of the sections also contains the items that were provided for consideration at the workshop to help kick off the discussion; these items were gathered through individual interviews, literature reviews, and discussions with the community at the First SEEDS Public Workshop.

7.1.1 Guiding Principles

Guiding principles are the high-level attributes of a reuse initiative and associated processes. A process that adheres to these principles should find acceptance within the Earth science community.

Guiding Principles Provided for Consideration

The following were provided for consideration at the second workshop as guiding principles for a reuse process:

- Defined by actual stakeholders in the ESE community
- Tailored to different environments rather than one-size-fits-all[38]
- Starts simple and evolves
- Ensures a practical focus
- Leverages existing activities & organizations
- Enables but does not force reuse
- Has cross-organizational emphasis…does not dictate individual organization practices
- Provides recommendations and guidance, but is not prescriptive

[37] Most of the community input comes from the Second SEEDS Public Workshop in San Diego, July 2002.

[38] Note that the community input indicates that the reuse *approach* should be different for each environment, but it is not clear at this time whether the process needed to support each approach is the same, varies slightly for each approach, or is substantially different for each approach.

Guiding Principles Identified by Mission-Critical Community

The community identified the following principles for mission-critical environments:

- Training and knowledge accompanies a software asset
- Different levels (granularity) of reuse can be appropriate
- New code developed should be written with reuse in mind; incentives to do this across projects are needed
- Use competition to drive reuse

Guiding Principles Identified by Mission-Success Community

The community identified the following principles for mission-success environments:

- Encourage reusability to be built into the original code (e.g., using team-programming approach, extreme programming) under the rationale that if you're programming it right the first time then sharing it is not a problem, and shared ownership of the code overcomes not-invented-here concerns
- Use defined criteria and community input to identify functional areas appropriate for reuse and likely to be successful, and focus reuse efforts on these areas
- Determine criteria for components to be included in an open source repository (e.g., community served, community size, interest of other communities in component & likelihood of reuse, structure of software)
- Create an open source authoring environment / infrastructure
- Determine the authority(ies) to modify open source software
- Include a cookbook for communities to be able to evaluate whether they should provide their software as open source
- Be scalable
- Define fast and streamlined approval process for proposed open source components
- Define peer-reviewed process for selecting components for open source

7.1.2 Contributing Factors

These are the positive and negative factors that contribute to the ability to reuse software and meet the goals of reuse (See section 0 "5.1 Evaluation Criteria."). As practical answers to the question of why reuse does or does not get done, a reuse process that addresses the contributing factors should have the characteristics needed to succeed in practice. Each factor is marked to indicate whether it has a positive (+) or negative (-) effect.

Contributing Factors Provided for Consideration

The following were provided for consideration at the second workshop:

- Intellectual property policies restrict sharing of reusable assets (-)
- Mission-oriented funding does not encourage development of assets for reuse (-)
- Developers of potentially reusable components do not have funding to document and support components for use by others (-)
- Access to experts (esp., component authors) reduces the effort and risk associated with reusing a component (+)
- "NIH", or the tendency to reject anything "not invented here" (-)

Contributing Factors Identified by Mission-Critical Community

The community identified the following opportunities (+) and problems (-) in mission-critical environments:

- Software modularity (+)
- Highly skilled workforce (+)
- Contractual relationships w/ vendors that limit access to software assets (-)
- Organizational bias to use certain software (-)
- Organizational knowledge about available software (+)
- Fit with application (+/-)
- Lack of contributions back to asset base (-)
- Lack of schedule/funds to make contributions to the asset base (-)
- Choice of language, esp. at beginning of mission may or may not match current preference (+/-)
- Changes to languages/technology (-)
- Proven code that does a hard or expensive function (+)
- Availability of test cases for reusable components (+)
- Lack of time at the beginning of a mission or on first of several missions (not practical to redevelop after the fact) (-)
- No reuse infrastructure (e.g., library of reusable assets, asset moderator, journal/forum for advertising assets) exists within the ESE community (-)
- Ability to start a small/simple process…get it out there! (+)
- Recognition that reuse could shorten schedule or get preliminary functions/foundation up quickly (+)
- No existing library of usable assets (-)
- NASA culture/mission (technical and management) encourages new tech prototypes over reuse (-)
- Intellectual property policies (across NASA centers and across contracts) restrict sharing of reusable assets.
- Mission-oriented funding does not encourage development of assets for reuse; developers of potentially reusable components do not have funding (and perhaps desire) to document and support components for use by others.
- Perceived technical risk (-)
- Cost impact on key resources to support reused components (-)

Contributing Factors Identified by Mission-Success Community

The community identified the following opportunities (+) and problems (-) in mission-success environments:

- Ability to assume some risk in development (+)
- Knowledge of--and communication with--those who have software assets (+)
- Hesitation to share because code was not designed/documented to be shared (e.g., TRMM IDL viewer) (-)
- Software modularity (+)
- Lack of incentive to reuse (e.g., no payment for support, mission-oriented funding, etc.) (-)
- Open source approach…inherently removes barriers, encourages design for reuse, improves quality, incorporates enhancements, results in feeling of shared ownership, etc. (+++)
- Licensing that requires contributions back to code base (+)
- Reuse approach established within a development team (+)
- Acceptance of limitations in existing SW (same as COTS) (+)
- Starting as open source vs. releasing to open source through lengthy export review (+)
- Issues regarding liability/copyright (-)
- Perceptions about security of open source- easy to find security holes (good if the good guys find them, bad if the bad guys find them) (+/-)
- Pride over authorship, control over functionality, and ability to match specific needs (-).
- Fear of having to support software released for reuse (-)
- Concern that effort to make it reusable will be wasted…asset might not be reused (-)
- Documentation not in consistent form (-)
- No ESE journal that supports advertising available assets (-)
- Perceived contribution toward interoperability (+)

7.1.3 Program Strategies

Program strategies are high-level plans of action that could be incorporated into a SEEDS management plan.

Program Strategies Provided for Consideration

The following were provided for consideration at the second workshop:

- Establish two working groups (mission-critical, mission-success) chartered with defining reuse processes and supporting architecture processes
- Establish a working group to develop program policies regarding reuse
- Establish teams to mine assets

- Establish forums to share reuse practices
- Enlist software engineering specialist support (e.g, CMU Software Engineering Institute)
- Enlist stakeholder representatives (or proxies) into the formulation study team until working groups can be established
- Top down vs. bottom-up: Start w/ architecture and work down to components vs. start with high-payoff functions and work up to system level
- In-place vs. external: have authoring organizations provide reuse support vs. have one or more non-mission orgs support reuse efforts

Program Strategies Identified by Mission-Critical Community

The community identified the following program strategies for mission-critical environments:

- For the mission-critical environment, start with a single working group
- Use working group(s) to set the policies/targets for level of reuse
- Minimize effort expended on infrastructure to enable reuse
- Have authoring organization provide technical support for reused components (for efficiency)
- Provide additional funding to cover overhead for initial activities needed to advertise and package reusable code (this is not to be confused with re-writing for reuse)
- Facilitate communication needed for reuse
- Do not enlist software engineering specialist support (CMU is overkill)
- Develop proper incentives to facilitate reuse

Program Strategies Identified by Mission-Success Community

The community identified the following program strategies for mission-success environments:

- Create an environment for conformance testing
- Start with prototypes to test/show/compare success
  - Leverage existing reuse-related resources (eg. Source Forge, Google) (potential issue with licensing)
  - Start with something small, simple, and general enough (to ensure broad interest) and see if it works
  - Try two prototypes: one involving outside communities and one involving ESE communities only
- Provide institutional support: create a program to encourage reuse
  - includes computers and staff
  - solicits software from community
  - polls community for validation of components
  - makes decisions about whether to support component

    - designs testbeds for testing component
    - identifies champion to oversee and support code
    - publishes what's available
- Reach out to other communities that may be interested only in specific components
    - Outside communities (in different domains) may share the need for specific components used in the ESE domain

7.1.4 Technical Strategies

A variety of technical strategies can be mixed and matched to enable software reuse.

Technical Strategies Provided for Consideration

The following were provided for consideration at the second workshop:

- Establish a reusable component library
- Establish a testbed to help identify and qualify reusable components
- Empower a team of experts to evangelize reuse
- Develop a software experience library with links to experts, assets, and other resources
- Develop a reference architecture to enable component reuse by enhancing component compatibility
- Document architecture/design patterns to enable design reuse
- Establish policies and incentives that counteract disincentives to reuse (e.g., NIH)
- Provides tools (e.g., reuse library software) to support reuse activities
- Incorporate reuse into NASA development standards
- Use formal methods and application generators to assemble systems from existing assets

Technical Strategies Identified by Mission-Critical Community

The community identified the following technical strategies for mission-critical environments:

- Develop software experience library with links to experts, assets, and other resources
- Do NOT focus on component library…an experience library is preferable
- Provide a standard checklist of items that must be provided to make assets reusable; refine the checklist periodically
- Do NOT establish a formal independent testbed…"testbedding" should be done at provider and customer site
- Conduct reuse workshops/forums to share lessons learned, advertise, obtain ideas, etc. at international meetings and other venues
- Document architecture/design patterns to enable design reuse

- Do NOT use formal methods and application generators to assemble systems from existing assets…this approach is overkill
- Clearly identify the originator of software to help indicate its quality

Program Strategies Identified by Mission-Success Community

The community identified the following technical strategies for mission-success environments:

- Host NASA open source workshops for awareness and training
- Collect and publicize success stories
- Clarify intellectual property issues
- Define management roles and responsibilities to encourage reuse (e.g., encourage team members to participate in open source forums)
- Provide incentives for sharing and reuse
- Identify and fund software to enable it to become successful open source (e.g., improving quality, going through the steps to provide it as open source, active engagement with community, moderating, etc)
- Examine other institutional models (HP, IBM, other governmental organizations, universities)
- Focus on reusable tools (with the rationale that scientists care more about the use of tools that work rather than the challenge of developing data systems)
- Focus on general software which has a broader opportunity for reuse as opposed to discipline and standard-specific software which has limited reuse opportunities
- Identify and focus on systems and components that have a high potential for reusability.

7.1.5 Evolutionary Approaches

Community recommendations and past experience both suggest that a reuse initiative will be more successful if it evolves over time.

Evolutionary Approaches Provided for Consideration

The following were provided for consideration at the second workshop:

- Strategy employed: enhancing communications → in-place-reuse → component library
- Specificity: Notional architecture → drill down to concrete architecture
- Formality: Prototype process → actual process
- Functional scope: highest-payoff functional areas/components → next highest areas

Evolutionary Approaches Identified by Mission-Critical Community

The community identified the following evolutionary approaches for mission-critical environments:

- A highly used module could, over time, be open-sourced
- End-user feedback to original provider (enhancements, bug fixes, etc.) acts like "mini tech infusion" and as an obsolescence fighter
- On-going participation of working group: more instances of reuse over time will help to map out the future directions

Evolutionary Approaches Identified by Mission-Success Community

The community identified the following evolutionary approaches for mission-success environments:

- Start with prototype process and learn from experience
- Identify and extend existing tools to support community needs (e.g., extend ENVI to support HDF)
- Start with high payoff and highly reusable areas

7.1.6 Reference Architecture Use

Reference Architecture Use Provided for Consideration

The following were provided for consideration at the second workshop:

- Enhance communications between groups that have software assets to share
- Standardize definition/functionality of components to help categorize components, enhance compatibility, and enable components to be independently acquired
- Make near-plug-and-play integration of components possible
- Enable black-box plug & play

Reference Architecture Use Identified by Community

The community identified the following reference architecture uses for mission-critical environments:

- Enhance communications between groups with software assets to share

The community identified the following reference architecture uses for mission-success environments:

- Enhance communications between groups with software assets to share

- Standardize definition/functionality of components to help categorize components, enhance compatibility, and enable components to be independently acquired
- As a guide, communication mechanism
- Document a set of requirements that a data system would meet (components needed, etc.)
- Enable reusing requirement specs
- Document the criteria for accepting components
- Create a market for components

7.1.7 Model Processes

The study team identified a number of activities related to reuse that could serve as models for a SEEDS reuse process:

- MODAPS/SeaWIFS/DODS development (various approaches)
- ESIP Federation IDL cluster (architecture-based and repository-based approach)
- Earth Science Modeling Framework (tool-based and architecture-based approach)
- National HPCC Software Exchange (tool-based approach)
- Reuse Information Clearinghouse (knowledge-based approach)
- National Association of State Chief Information Officers ComponentSource (repository-based approach)
- Workshop on Institutionalizing Software Reuse (knowledge-based approach)

At this time, these activities have not been assessed by the community or study team to determine if any should be used as a model for a SEEDS reuse process, but they are noted here for future reference.

7.1.8 Highlights of Community Input and Recommendations

The study team attempted to capture all community input in the preceding sections. However, judging by how often various recommendations were raised and the level of agreement expressed by workshop participants, some items are clearly more important and more broadly accepted than others. They are summarized below for emphasis.

- Do something, because reuse across projects rarely happens by itself. In many cases, only a small amount of additional effort or funding may be needed to make valuable software available outside the group that developed it. In other cases, significant barriers (especially intellectual property policies) may have to be removed or circumvented.
- Start with a simple process and engage all stakeholders in refining and evolving it. Leverage existing resources, infrastructure, forums, etc. as well as lessons-learned from similar initiatives to ensure that real results are achieved quickly and cost effectively.

- Use competition and peer-review rather than blanket policies to drive reuse to help ensure that reuse always serves the ESE goals and does not become an end in itself.
- A record of authorship and access to authors is essential for an asset to be reused. In this regard, code is treated much like science data: usage often boils down to quality and trust in the source.

7.2 Notional Reuse Process

7.2.1 Summarization of Community Input

In keeping with the principles identified during the study, the effort to define a reuse process will continue to be community based. To provide a point of departure for future discussions on this topic, the study team synthesized a notional process based on the community input contained in the sections above. The suggestions here are only intended to "bootstrap" the process, which could change significantly based on community input.

A substantial amount of guidance from the community relative to a reuse process has been captured in Section 0 "Reuse Process and Next Steps". To summarize, input from the community indicated the following:

- **Principles.** A SEEDS reuse process should be a community-owned, non-prescriptive, scalable, practical process that starts simply and evolves, emphasizes directly enabling reuse over infrastructure activities, and relies on competition and peer review rather than mandates to drive reuse appropriately.
- **Contributing Factors.** There are currently a number of barriers to reuse, including project funding constraints, licensing issues, support concerns, security concerns, cultural issues, and communication issues. A process that removes some of these barriers could significantly improve the level of reuse within the ESE.
- **Program Strategies.** A reuse process should emphasize community-based working groups, fund actual reuse activities rather than infrastructure activities, and establish incentives to encourage reuse. To a lesser extent, a reuse process should facilitate sharing of knowledge related to reuse and reusable assets, provide some institutional support, and establish/revise policies to further enable reuse. Again, the principle of starting with a small, simple process and building on what works was emphasized.
- **Technical Strategies.** A reuse process should employ a variety of technical strategies including information sharing (through workshops, contact directories, published success stories and best practices, checklists, etc.), quality indicators (e.g., identifying component authors), and direct funding (e.g., of documentation/generalization/support for components used across projects). At the level of technical strategies, differences between environments become more apparent. For example, the mission-critical / improved clone and own group favored in-place support from authoring organizations and no component library, while the mission-success / open source group favored establishing an open

source infrastructure utilizing existing tools. The input indicates that technical strategies focused on methodology, policy enforcement, and automatic programming are not appropriate. The mission-success / open source group emphasized that, regardless of the technical strategies employed, it is important to focus on components with a high likelihood of reuse.

- **Evolution.** The process should evolve primarily in terms of the definition of the process itself (starting simply and learning from experience), and also in terms of focus (i.e., which functional areas have the highest potential payback).
- **Reference Architecture Use.** It seems clear that the community is interested in knowing what components are available to meet a specific need, and that a reference architecture should, above all else, provide the definitions needed to ensure effective communications between component suppliers and component users.

As the notional process is refined into a working process and then evolved, it should continue to be validated against community input. In addition, relevant critical success factors derived from industry experience should be identified and used to guide the process. A good starting point is the work of Morisio[39], Rine[40], and Tracz[41], as well as the recommended practices from the proposed SEI CMM Software Reuse Key Process Area[42].

Particularly important to the SEEDS formulation study at this time is the community input regarding program strategies. The possible program strategies boil down to six basic options:

- **Charter one or more groups to define and support a reuse process.** This strategy was uniformly recommended by the community representatives that attended the SEEDS workshops. In keeping with community recommendations, the chartered groups need to be drawn from the ESE community rather than enlisting the support of outside experts. To ensure the integrity of the process, the chartered groups should be “working groups” that make technical decisions and funding recommendations, with all significant funding decisions remaining with NASA management. To avoid conflicting goals and priorities, the study team recommends establishing separate working groups for each environment (i.e., mission-critical and mission-success). For a quick start, stakeholder representatives/proxies could be enlisted into formulation study teams until the working groups can be established.

[39] M. Morisio, “Success and Failure Factors in Software Reuse.” IEEE Transactions on Software Engineering, Vol. 28 No. 4, Apr 2002.

[40] David C. Rine and Robert M. Sonnemann, “Investments in reusable software. a study of software reuse investment success factors”. Journal of Systems and Software, 1998.

[41] Will Tracz, *Confessions of a Used Program Salesman: Institutionalizing Software Reuse*. Addison-Wesley, 1995.

[42] This proposed KPA was not accepted into the SEI CMM, but provides a concise reference for recommended reuse practices.

- **Fund actual software reuse efforts.** By "actual software reuse efforts" we mean activities primarily concerned with taking a reusable asset and employing it in an operational system, including identifying, qualifying, and adapting reusable assets. This strategy was uniformly recommended by the community. In keeping with community recommendations, authoring organizations should be funded to provide technical support for reused components because it is more efficient than paying other groups to learn components well enough to support them. Authoring organizations could also be funded to package, document, and advertise their software assets. Within this strategy, peer-review or a competitive process should be used to allocate funding. At the beginning, SEEDS should fund trial efforts and review the outcome to validate the expected return on investment and help refine the process before more significant investments are made.
- **Establish incentives to encourage reuse.** This strategy seems uniformly recommended by the community with the caveat that the incentives should only *enable* (rather than *promote*) reuse by helping to overcome artificial barriers to reuse.
- **Facilitate sharing of knowledge about reuse and reusable assets.** This strategy seems supported by the community, though more as an implied solution to known problems rather than as a direct recommendation. Specific recommendations include publishing available components, sharing success stories, and conducting reuse sessions at related ESE conferences.
- **Fund activities that support/enable software reuse.** This strategy has mixed support within the community. The community recommended forming testbeds by using or linking existing resources to provide a means of validating reusable components. However, most of the community representatives recommended against making significant investments in infrastructure to support reuse, preferring instead to focus on direct reuse efforts. Specifically, those in mission-critical environments recommended against building a reusable component library, and those in mission-success environments thought that existing tools and services could be employed to meet the needs of open source component development and reuse. Still, modest efforts to develop a notional reference architecture to help characterize reusable components, to implement an open source software development environment from existing tools, and other activities that help remove some of the identified barriers could make a substantial contribution to reuse.
- **Establish/revise policies related to reuse.** This strategy has support within the community with certain qualifications. In keeping with community recommendations, the focus should be on *enabling* reuse by working to change policies that currently inhibit it. For example, specific concerns were raised regarding NASA and university policies in the area of intellectual property rights. Those in the mission-critical environment recommended that the responsibility for developing policies and targets for reuse be delegated to a community-based reuse working group. It is important to note that the community recommended against using policies (e.g., contract or grant language) to *drive* reuse because such an approach could require reuse in circumstances where it is not appropriate.

These program strategies can serve as the initial framework for a notional reuse process to define and conduct a set of small reuse initiatives.   The following section incorporates these strategies into a notional process as viewed from a program management perspective.

7.2.2 Program Management Perspective

The following diagram depicts how specific reuse initiatives could be identified and pursued over time. It is important to emphasize that this notional process is based onthe community input summarized above, but has not itself been subject to community review. The key elements of this process are a set of small ESE Reuse Initiatives that are implemented through a variety of reuse projects and activities; ESE Community Reuse Implementers who actually perform reuse implementation projects; a SEEDS Integration Office that (among its other duties) is responsible for the reuse initiatives; and community-based ESE Reuse Working Groups that perform certain reuse activities. .}

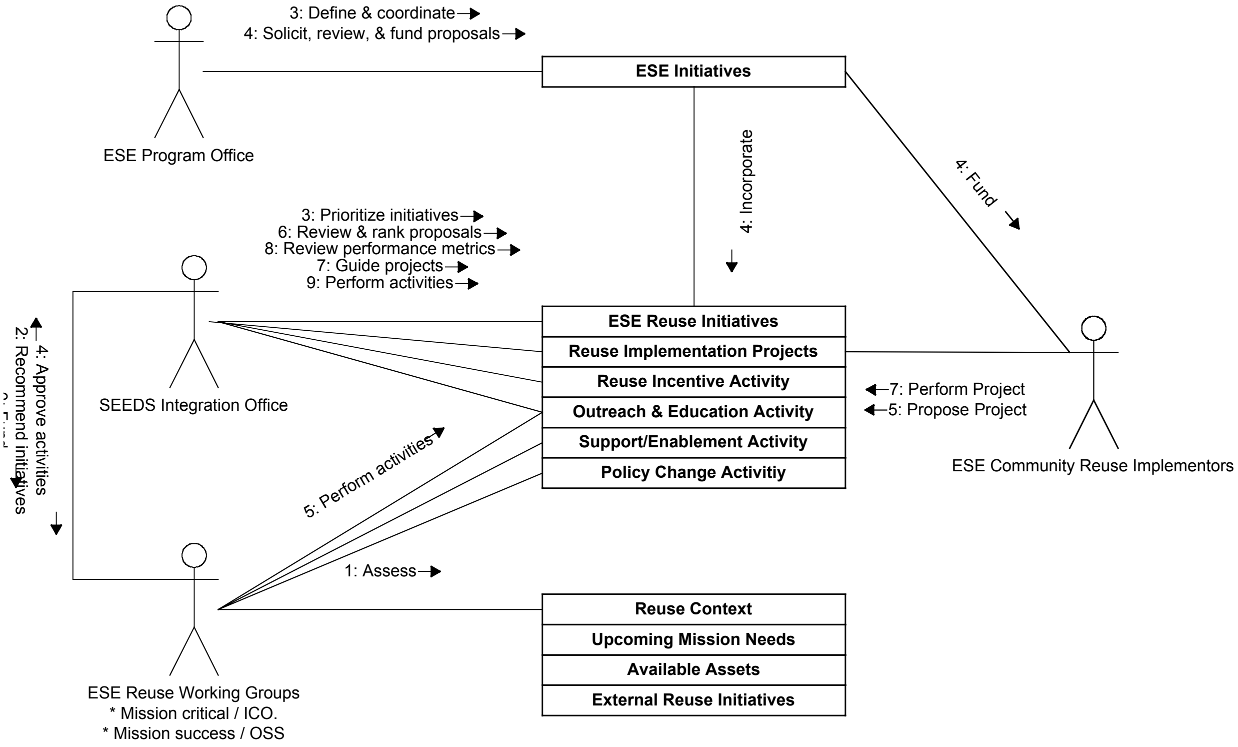


Figure 0-1. A notional reuse process that incorporates the input from the ESE community on process principles, program strategies, technical strategies, evolutionary strategies, and reference architecture use.

In this notional process, the SEEDS Integration Office funds two ESE Reuse Working Groups: one focused on the improved clone and own reuse approach for mission-critical environments, and one focused on the open source reuse approach for mission-success environments. The working groups are small teams composed of representatives from the ESE community, thus ensuring deep community involvement in all reuse initiatives.

Working groups are competitively selected through the same mechanisms used to select reuse initiatives (e.g., CANs).

The Working Groups are responsible for recommending ESE reuse initiatives to the Integration Office and performing certain approved activities. Specific tasks include reviewing the needs of upcoming missions, comparing these needs against the current catalog of assets and external reuse initiatives to identify gaps, identifying candidate initiatives to fill the gaps, and using the reference architecture to identify and prioritize synergistic activities. They further qualify candidate initiatives against the SEEDS program goals (e.g., initiatives that have the highest potential for reducing the cost of ESE data systems; increasing flexibility and responsiveness to future missions, science, and applications; and increasing effective and accountable community participation).

The Integration Office prioritizes the candidate initiatives recommended by the Working Groups, works with the ESE Program Office to incorporate these initiatives into various solicitations, and participates in reviewing and ranking proposals submitted by ESE Community Reuse Implementers. After projects are awarded, the Integration Office provides advice and guidance to help projects focus on high priority areas and leverage the work of other related projects. As projects provide results, the Integration Office reviews performance metrics to determine the success not only of individual projects, but also the performance of each overall initiative and the total reuse effort. The separation of responsibilities between the program office and working groups allows interested members of the ESE community to participate fully in the process without concern of being excluded from projects because of conflict-of-interest concerns.

ESE Reuse Initiatives with specific goals (e.g., to make reusable assets available or better publicize reusable assets) that may be accomplished through a variety of activities. The five activities identified in the diagram above are based on the summary list of program strategies, and are listed roughly in order of emphasis based on community input. **Reuse implementation projects** directly result in the publication or use of a reusable component. Examples might include component generalization and documentation, reusable component integration pilots, and component qualification, although the actual projects are determined by proposals from the ESE Community Reuse Implementers. **Reuse incentive activities** include awards and structural changes that indirectly encourage reuse. Examples might include small competitive or lottery-type monetary awards (for submitting components to a repository, authoring a popular component, etc.) and reuse incentives tied to technology programs (e.g., funds saved through reuse could be spent on further R&D). **Outreach and education activities** increase reuse by increasing the ESE community's understanding of the benefits, best practices, tools, etc. relating to reuse, and increasing awareness of available components. Examples might include conference workshops, contact directories, Web sites, newsletters, and articles in Earth science journals. **Support/enablement activities** increase reuse by providing tools and mechanisms to enable reuse. The most important example is developing a reference architecture to facilitate communications between component suppliers and potential component users. Other examples might include implementing (not developing) asset catalogs and open source software development Web sites, providing open source license templates, and linking existing testbeds for demonstrations. **Policy change activities** are intended to reduce policy barriers to reuse, such as revising open source software policies.

The Integration Office performs work in two areas: Reuse incentive activities, and outreach and education activities. Specifically, the Integration Office administers incentives and works to provide visibility to ESE reuse activities, best practices, and success stories. The Integration Office includes a small amount of technical support with expertise in reuse to help conduct these activities, as well as to provide coordination across the other projects and activities of the Working Groups and ESE Community Reuse Implementers.

The Working Groups perform work in three areas: Outreach and education activities, support/enablement activities, and policy change activities. Specifically, the working group is responsible for developing the reference architecture.

7.3 Next Steps

An ESE reuse implementation timeline cannot be established until such an initiative has been approved and a process has been defined. A notional timeline of activities by year starting roughly in FY2003 follows:

1. Groundwork. Begin to engage the Earth science community via workshops, surveys, and consultations. Conduct some analysis of reuse costs, benefits, issues, and processes. Begin to define implementation plans (including staffing, tools, etc.) and required budgets. Examine policies for linking reuse with technology infusion and science capability enhancement. Examine intellectual property policy issues. Identify high-payback reuse areas/components (possibly a data server or user interface widgets. Define an implementation/management plan.
2. Initiation. Begin to establish a reuse infrastructure. Begin development of a notional architecture. Create a "paper prototype" of a reusable asset collection focused on one of the high payback areas. Issue AO for reuse testbeds.
3. Execution. Maintain and operate reuse libraries and other infrastructure. Support new projects with expertise from existing projects. Develop concrete architecture for selected components. Populate reuse library: extract assets, assess compliance with architecture using testbeds, modify to improve reusability. Capture and publish key artifacts (architecture/design documents, test plans/data, etc.).
4. Iteration. Capture and disseminate additional software assets. Revise reference architecture and make more concrete in functional areas as needed. Issue AO for additional testbeds.

8 Additional Information

The slides used at the SEEDS workshop to introduce and conduct the evaluation of reuse and reference architecture alternatives can be found on the SEEDS public Web site, http://lennier.gsfc.nasa.gov/seeds/.